UNPACKING THE BLACK BOX OF ICO WHITE PAPERS: A TOPIC MODELING APPROACH

Anna M. Pastwa, Prabal Shrestha, James Thewissen, Wouter Torsin

LIDAM Discussion Paper LFIN
2021 / 18

UCLouvain

LIDAM
LFIN | Louvain Institute of Data Analysis and Modeling in economics and statistics

LFIN

Voie du Roman Pays 34, L1.03.01 B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/lfin/publications.html

Unpacking the black box of ICO white papers: A topic modeling approach

James Thewissen
Université catholique de Louvain
james.thewissen@uclouvain.be

Prabal Shrestha
Katholiek Universiteit Leuven
Université catholique de Louvain
prabal.shrestha@kuleuven.be

Wouter Torsin
HEC Liège
wtorsin@uliege.be

Anna M. Pastwa
Warsaw University of Technology
pastwa.anna@gmail.com

ABSTRACT

We apply a novel topic modeling method to map Initial Coin Offerings' (ICOs') white paper thematic content to analyze its information value to investors. Using a sentence-based topic modeling algorithm, we determine and empirically quantify 30 topics in an extensive collection of 5,210 ICO white papers between 2015 and 2021. We find that the algorithm produces a semantically meaningful set of topics, which significantly improves the model performance in identifying successful projects. The most value-relevant topics concern the technical features of the ICO. However, we find that white paper's informativeness substantially diminishes after the token is listed. Moreover, we show that credibility-enhancing mechanisms (i.e., regulations and ICO analysts) reinforce the information value of ICO white papers. Overall, our results suggest that the topics discussed in white papers and the attention devoted to each topic are useful ICO performance indicators.

***JEL* Classification:** G15, M13, L26, D80

Keywords: ICOs; White paper informativeness; Topic modeling; ICO regulation; ICO analysts.

Declaration of Interest: None

Corresponding Author: James Thewissen, LIDAM Voie du Roman Pays 34 / L1.03.01, 1348 Louvain-la-Neuve, Belgium. james.thewissen@uclouvain.be

1. Introduction

Initial Coin Offerings (henceforth, ICOs) provide blockchain entrepreneurs with the opportunity to raise external financing from investors at an early stage of their venture without a financial intermediary. During an ICO, investors purchase issued coins (or 'tokens'), which can be traded with other investors or exchanged for the firms' goods or services. Despite an impressive growth, the lack of regulatory enforcement from public institutions and the severe level of information asymmetry remain ongoing hurdles for the development of this market (Tiwari, Gepp, and Kumar, 2019).[1] Compared to the traditional IPO market, there are distinct informational challenges as investors base their investment decisions on scarce, unaudited and voluntarily disclosed information. In fact, the white paper is the main document that investors use while making their capital allocation decisions in an ICO (Chod and Lyandres, 2021; Fisch, Masiak, Vismara, and Block, 2019). Often likened to the prospectus from private firms going public, the white paper predominantly pitches the project idea, detailing the business plan, the team and the underlying risks (Florysiak and Schandlbauer, 2021). Despite the importance investors attach to the white paper (Fisch et al., 2019; Lyandres, Palazzo, and Rabetti, 2020), there has been no systematic effort to examine its actual content. By relying on unsupervised machine learning, this study investigates white papers' thematic content and its informativeness in explaining the ICO outcome and post-ICO performance.

Given the prominence of white papers in investors' decision-making, it is essential to identify the components that help influence investors' decisions. Recently, a growing body of research examines whether narratives and the associated language in white papers may help leverage resources by conveying a comprehensible identity for an entrepreneurial firm, with mixed results. For instance, Fisch (2019) and Bourveau, De George, Ellahie, and Macciocchi (2021) show that more technical white papers are a useful signal to investors and can help predict ICO success. Zhang, Aerts, Lu, and Pan (2019), Samieifar and Baur (2020), Feng, Li, Wong, and Zhang (2019) and Dittmar and Wu (2019) further document that the textual style characteristics, such as readability, exaggeration, or tone of the narrative, can help identify successful ICOs. In contrast, Adhami, Giudici, and Martinazzi (2018) observe no significant relationship between the availability of white papers on the probability of reaching the stated funding goal. In addition, Momtaz (2020a) observes only a marginal relationship

[1] Note that the lack of regulatory oversight remains a salient feature of the ICO process, the ramifications of which are worsened by the anonymous character of token transactions. Blaseg (2018) reports that ICOs that impose investor restrictions such as whitelists or Know-Your-Customer (KYC) requirements to address such challenges face a reduced pool of potential investors and longer fundraising time.

between the number of words in white papers and ICO success and post-ICO performance. Similarly, Florysiak and Schandlbauer (2021) decompose the content of white papers into informative and standard contents based on disclosure norms and find no direct relationship between informative content and the amount raised. Against this backdrop, this study aims at unpacking the black box that constitute white papers and investigates the information value of white papers by examining *what* is contained in white papers instead of *how* information is communicated.

We propose that the thematic composition of white papers contains information value to explain ICO success. The information content of white papers, such as tone, the number of words, or readability, has frequently been highlighted as a key determinant of the success of ICOs (Zhang et al., 2019; Sapkota and Grobys, 2021). However, if word count or length may capture the amount of information contained in a white paper, it is a weak proxy, if at all, for the quality and type of information conveyed in the document. In fact, such linguistic attributes ignore the underlying meaning or context of the disclosure and consider words as independent and informative units, thereby limiting the inferences that can be drawn (Loughran and McDonald, 2016). In addition, using word counts to measure the informativeness of ICO white papers needs further reflection given the level of heterogeneity and the lack of a formal control mechanism or standardization. As argued by Hoberg and Lewis (2017), the flexible nature of disclosure content requires a more extensive set of dimensions along which we could analyze white paper narratives. Therefore, there is an urgent need to examine the content of white papers and demonstrate methods that accommodate contextual subtleties and informational differences across documents. In light of this background, in this study, we explore three research questions: (i) what are the topics discussed in ICO white papers, (ii) which topics are associated with ICO funding success, and (iii) whether thematic content is informative in explaining post-ICO performance.

The thematic content of a white paper is defined as the distribution of underlying topics in a document, and is measured as the number of sentences relating to a specific theme. To identify the topics included in white papers, we use a novel process-based approach based on the Latent Dirichlet Allocation (LDA) at the sentence-level (sentLDA) introduced by Bao and Datta (2014). This method reasonably assumes that a sentence is the smallest integral unit of text that conveys a complete and meaningful idea (Ivers, 2010), and thus incorporates the information in sentence boundaries while identifying the topic clusters (see Bao and Datta, 2014). Consequently, it allows us to identify a topic for each sentence instead of solely estimating the topic distribution of the entire document. Because

the purpose of our paper is to identify and quantify the white papers' content to understand its information value, rather than simply classifying the documents, sentLDA is particularly suited for the task. The identified topical composition is then used to discover cues affecting the investment decision of ICO investors. Correspondingly, an investigation of the relationship between the discovered topics and ICO success and post-ICO performance is included as a part of the study.

We map the informational composition of 5,210 white papers for coin offerings that ended between August 2015 and June 2020. We find 30 optimal topic categories and evaluate their semantic validity using both human and machine-based procedures.[2] Altogether, we find that the sentLDA method produces a coherent set of meaningful topics that capture ICO white papers' information content. Our results also show significant diversity in the topics covered in ICO white papers, ranging from technical descriptions, such as underlying blockchain structure, smart contracts, and data protection, to business-related concerns, such as future roadmap, market size, and risks. We find that blockchain application is the most discussed topic in the analyzed collection, followed by information on the network's development and discussions regarding data management and the application of artificial intelligence tools. Apart from the emphasis on blockchain technology, we also observe that ICO white papers distinctly entail substantial discussions on decentralization and network building, energy consumption and sustainability, and topics concerning industries, such as the financial sector, health and gaming. Topics related to legal disclaimers, risk management and risk disclosures appear to be the least discussed information in white papers.

We next investigate the value-relevance of each topic for investors to identify successful ICOs. To do this, we examine the relationship between white papers' thematic content and the likelihood of token issuance, the amount raised, and the time-to-token-listing. Our tests on out-of-sample model performance indicate that including the thematic content significantly improves the models' capacity to explain successful ICOs relative to models that include standard quantitative and text-based control variables used in prior literature. We find that several topics are consistently associated with successful ICOs. A key highlight from our findings is that successful ICOs devote significantly more attention to technical details concerning blockchain and mining. Overall, these results indicate that investors use white papers to allocate their capital among the set of ICOs. However, more importantly, they also show that ICO white papers' thematic content contains information value that is incremental to the variables identified in prior literature.

[2]We simulated topic distributions with several alternatives (15, 20, 25, 30, 35, 40 and 50 topics). We find that 30 topics capture the most meaningful topic categories while preserving sufficient distinctiveness among topics.

In contrast, we find that the informativeness of the white papers' content substantially diminishes after the token is listed. While we do find some relevance of white paper information in relation to first-month token-price volatility and the probability of eventual delisting, the overall association with white paper topics is notably weaker. For first-day return and six-month long-term return models, not more than a single topic category is significant: (i) topics relating to security are negatively and significantly correlated with the ICO's initial returns and volatility, (ii) the long term returns of ICOs are positively associated the topic of token security, while (iii) ICOs that discuss security more tend to remain listed on the market longer. Other topics, such as ICO characteristics, blockchain, network, or security that were significant in explaining ICO outcome, are no longer consequential. Our results suggest that, as the project gains recognition and investors have access to other sources of information, the influence of information in ICO white papers subsides in the post-ICO period. These findings concur with the efficient market hypothesis and suggest that, after a successful ICO, supply and demand dynamics start to form an equilibrium price based on all available information about the firm. The token's exchange listing then performs a coordination function that dispersed investors fail to undertake on their own during the ICO (Momtaz, 2020a). Therefore, post-listing, information asymmetry decreases as pricing information becomes available, making the content of the white papers less relevant.

Apart from the topics' inherent attribute in explaining ICO performance, the degree to which voluntary disclosures, such as white papers, mitigate resource misallocation is likely to depend on external factors that lend credibility to the disclosure. In the second part of this paper, we therefore focus on the impact of two credibility-enhancing mechanisms that may influence the informativeness of ICO white papers. First, given that ICOs remain a novel market with fast-evolving regulations, we examine the association between the host country's regulatory status on the informativeness of the white paper's thematic content. To test this relationship, we identify a country's regulatory status regarding ICOs based on the data of the study of Shrestha, Arslan-Ayaydin, Thewissen, and Torsin (2021). We find the white paper to be more credible for ICOs from countries with specific regulation. In fact, there exists a significant contrast in the information value of ICO white papers in countries with and without ICO-specific regulation. In regulated countries, themes in white papers significantly explain ICO outcome, while only a limited selection of topics explain the post-ICO performance. On the contrary, in countries with limited regulations surrounding ICOs, the white papers' thematic content only marginally relates to the ICO outcome, but has a more significant association with

post-ICO performance. We attribute this difference to the trust differential between investments in ICOs from regulated and unregulated countries (Shrestha et al., 2021). When asymmetric information is at play during the token offering, investors exhibit low trust levels in white papers from projects based in unregulated countries and subsequently discount white paper information in their investment decisions. However, once the token is listed, the ICOs from unregulated countries receive a stamp of approval from the market, and investors start to incorporate the information contained in the white paper.

Second, we examine the impact of newly emerged information intermediaries, such as ICO analysts, on the information value of ICO white papers. ICO analysts are experts who voluntarily provide ratings on the quality of ICOs to rating platforms (see e.g., Lee, Li, and Shin, 2021; Bourveau et al., 2021; Barth, Laturnus, Mansouri, and Wagner, 2021). Given that these expert ratings bridge the informational gap between investors and ICO issuing firms, we examine whether favorable ratings enhance the eventual relevance of ICO white papers. We observe notable differences in the influence of white paper thematic content in terms of ICO success between high- and low-rated projects. Consistent with our initial results, we find that the thematic content of white papers from high-rated ICOs is significantly associated with ICOs' success. We even find that good ratings improve, to a degree, white papers' association with post-ICO's performance indicators, namely initial volatility, long term returns and the probability of being delisted. However, among low-rated ICOs, we find that the white papers' informativeness is substantially reduced, for both ICO success and post-ICO performance. Overall, our results complement the findings of Lee et al. (2021), and Bourveau et al. (2021) and highlight the positive certification role played by information intermediaries, such as ICO analysts, in a market with severe informational constraints.

Our paper contributes to the understanding of the ICO market in several ways. First, based on a comprehensive sample of ICOs, our study adds to the nascent research on ICOs, which is yet to offer adequate evidence on the role played by white papers' content on ICOs' success. In a survey examining ICO investors' motivations, Fisch et al. (2019) find that while only a small minority (3.1%) of respondents state that they generally do not read white papers, 31.5% indicate that they 'read the white paper in detail and try to understand everything.' Despite investors' interest in white papers, there has however been little research on which specific information in white papers is associated with the ICO's performance. In fact, most studies examining the white papers' role in ICOs tend to ignore the context or meaning within these documents, and overall are limited to the study of stylistic

attributes (Zhang et al., 2019; Zhang, Aerts, Zhang, and Chen, 2021; Samieifar and Baur, 2020; Feng et al., 2019; Dittmar and Wu, 2019; Sapkota and Grobys, 2021). With this study, we move beyond the restrictions in prior literature. We are the first to provide an in-depth descriptive analysis of white papers' thematic content and to evidence that, despite the voluntary nature of these documents, the thematic content of ICO white papers is informative of the ICO's outcome, but not of the post-ICO performance.[3] Although our results are descriptive and illustrate associations instead of causal relationship, our results enrich the discussion on the information cues important to ICO investors (see e.g., Florysiak and Schandlbauer, 2021; Fisch, 2019; Samieifar and Baur, 2020) by providing evidence on which topics are associated with successful ICOs (e.g., blockchain- and network-related information) and the supporting role played by white papers' thematic content for firms in th e ICO market.

Second, our work finds itself at the crossroads between artificial intelligence and finance and is the first to exploit a robust machine learning tool at the sentence level that quantifies *what* is being disclosed in ICO white papers as opposed to *how*. This content analysis tool is a significant step forward, as it delves deeper into how one can use automated tools to extract meaning from such unstructured documents and quantify their informational content without relying on *a priori* assumptions. This approach is distinct from that of recent studies, such as Bourveau et al. (2021) and Fisch (2019), which use hand-collected pre-identified disclosure items in white papers. Despite the importance of their contributions, such an *ex ante* identification of what white papers may disclose based on the researchers' judgment is not suitable to construct a detailed map of the document's informational content. Moreover, relying on human-coded indices has limitations, including that of scalability and inconsistency, especially in dealing with large heterogeneous unstructured textual data such as white papers (for related discussions, see Lewis and Young, 2019; El-Haj, Rayson, Walker, Young, and Simaki, 2019). In addition, our method significantly departs from other recent studies that use computer-based tools to operationalize informational components within white papers. For instance, Florysiak and Schandlbauer (2021) decompose white papers' text into informative and standard contents based on term frequencies within documents, quantifying informativeness as deviations from industry-wide and recent disclosure norms, while Lyandres et al. (2020) use a bag-of-words approach on a sample of over

[3]One exception applies, however. Fu, Koh, and Griffin (2019) use the traditional LDA method in order to classify ICOs into different industry categories. Their results provide an interesting application of unsupervised machine learning-based topic modeling to classify ICOs. However, our objective significantly departs from their paper as we do not intend to classify ICOs, but aim at leveraging the outcome of the (sent-)LDA method to quantify the information value of white papers and identify the topics of value-relevance for investors.

1,100 white papers to show that the use of unique and technical words in white papers is positively associated with ICO success. However, instead of focusing on pre-defined information categories or human coders, our study relies on an unsupervised method, e.g. topic modeling, to discover the informational content within each white paper, mitigating the influence of researcher's *ex ante* judgments while mapping the detailed information composition in issuing firms' communications. This method will not only highlight the large heterogeneity of themes covered in white papers, but also allows us to leverage this information to identify the topics associated with successful ICOs. As the ICO market and the data around it expand, the need for such computer-aided tools, which are virtually unconstrained by processing limitations or subjective biases, will only grow. The use of the sentLDA method introduced by Bao and Datta (2014) is one step in that direction, and responds to the call by El-Haj et al. (2019) for 'new horizons in textual analysis.'

Finally, we contribute to the debate on the role of voluntary disclosures in the ICO market (Botosan, 1997; Leuz and Verrecchia, 2000; Lee et al., 2021; Healy and Palepu, 2001; Bourveau et al., 2021) and highlight the importance of credibility-enhancing mechanisms such as external monitoring or regulation in supporting the informativeness of voluntary disclosures. Although the ICO market's rapid rise has led to substantial scrutiny from analysts and authorities around the globe, regulation on ICOs remains lax in many jurisdictions, where authorities are taking a "wait-and-see" approach to better understand the implications of this novel fundraising method (Tiwari et al., 2019). However, as discussed theoretically by Chod and Lyandres (2021), it has become clear that, for ICOs to remain a legitimate alternative for financing entrepreneurial ventures, they should be regulated in some way. Based on the results and methodological approach presented in this paper, regulators can draw a better understanding of what type of information in white papers decrease information asymmetry and propose a standardization framework for white papers' content structure that mitigates the issues of asymmetric information. We also inform regulators that, in the absence of regulation, new types of information intermediaries such as ICO analysts can increase the credibility of ICO white papers' content and improve the market's quality.

2. Background and research questions

2.1. Initial Coin Offerings

ICOs are a form of venture financing involving a crowd-sale of tokens to investors, which ultimately allow firms to tap into international capital at a low cost and minimal regulatory restrictions (for excellent introductions to ICOs, we refer the reader to Howell, Niessner, and Yermack (2020), Momtaz (2019), and Amsden and Schweizer (2018)). Distinct from traditional shares, tokens can resemble various financial instruments, including debt, financial derivatives, or even a right to future goods or services. These obligations are enforced autonomously through smart contracts without any mediating institution. Furthermore, the issued tokens are liquid, readily tradeable in a secondary market with few obstacles in transactions across national borders (Bakos and Halaburda, 2018; Benedetti and Kostovetsky, 2021). The number of ICOs exploded in 2017. In fact, the cumulative funding amount already exceeds the entire European venture capital industry, and the largest token offering so far (EOS, $4.2 billion) is comparable in size to the three largest IPOs in the same period. It also exceeds the cumulative funding amount of all crowdfunding initiatives of the premier platform (Kickstarter) since its inception in 2009, as summarized by Momtaz (2020a).

Typically, an ICO process starts with the publication of its white paper. This document, made available on the project's website and ICO listing sites, describes the proposed undertaking (Florysiak and Schandlbauer, 2021). The firm uses this document to present and promote her project to potential investors, detailing the project's value proposition, technical features, team, background, and objectives (Tasca, Cerchiello, and Toma, 2019). Apart from the details provided in these white papers, reliable information for investors is scarce. Investors cannot rely on detailed due diligence, which is common in venture capital or angel transactions. In addition, unlike crowdfunding, there is no mediating platform with an inherent incentive to weed out bad actors (Agrawal, Catalini, and Goldfarb, 2014). This lack of reliable and credible information is amplified by the fact that there is little direct access to issuers, who are predominantly early-stage ventures without proven track records and developed products (Fisch, 2019). Furthermore, any regulatory action to ensure reliable disclosure is challenging to implement, as issuing firms are not obliged to associate with any legal jurisdiction (Howell et al., 2020).

Given its distinguishing attributes, this nascent financing innovation has spurred a growing body of research on ICOs, a significant portion dedicated to understanding the factors that influence ICO

success and post-funding outcomes. For instance, research shows that ICOs that involve a pre-sale and bonus schemes, tokens based on a new blockchain protocol, and tokens linked with a real asset are more likely to be successful (e.g., Adhami et al., 2018; Roosenboom, van der Kolk, and de Jong, 2020; Lyandres et al., 2020). Manager-specific attributes, such as connectedness and CEO loyalty (Amsden and Schweizer, 2018; Benedetti and Kostovetsky, 2021; Momtaz, 2020b), and the firm-specific features, such as team-size and country-of-origin, are also linked with funding success and post-ICO outcomes (Amsden and Schweizer, 2018; Shrestha et al., 2021). In addition, as ICO funding lacks an overseeing intermediary, the question of how investors respond to firms' unverified disclosures remains a compelling question. From recent studies, we know that communications with greater transparency and technical details encourage investors and indicate the firm's future performance (Howell et al., 2020; Roosenboom et al., 2020; Fisch, 2019). Our study builds on this latter stream and examines whether and how the white paper's informational content relates to the ICO's performance.

2.2. ICO informational context and the role of white papers

The ICO market's potential to become one for lemons becomes apparent in the model of Chod and Lyandres (2021). The severe level of information asymmetry is structural, mostly as the supporting blockchain supplants the need for a mediating third-party. Yet, a predominant ideal permeating through the industry and crypto-investors is the minimal reliance on intermediaries or governmental supervision (Chen and Bellavitis, 2020). As indicated by Kim, Miller, Wan, and Wang (2016), intermediaries play a crucial role in incentivizing information generation, which mitigates the problem of asymmetric information between investors and issuing firms (Benveniste and Spindt, 1989). However, trusted institutions, either public or private, that could improve the informational challenges remain mostly absent from the ICO market (Momtaz, 2020a).

Because of the lack of monitoring, there is no disclosure requirement on token sales. The publication of a white paper is a voluntarily practice without any legal or regulatory obligation (Amsden and Schweizer, 2018). Early studies by Howell et al. (2020) and Adhami et al. (2018) find that only around 80% of the issuers published a white paper before the ICO. More recently, in a study based on token offerings before November 2018, Momtaz (2020a) obtain the white papers for less than 50% of the ICOs in their sample. Furthermore, there is no standardized disclosure format dictating what a white paper should or should not include. In certain instances, a white paper is a one-page document containing basic financial information on the token sale. Most ICO white papers are found to include

little information that would help assuage investors' concerns about insider self-dealing (Cohney, Hoffman, Sklaroff, and Wishnick, 2019). In fact, studies such as Zetzsche, Buckley, Arner, and Föhr (2017) find that many white papers even lack essential information such as contact information or the names of the individuals behind the project.[4] However, the authors also find that several white papers were professionally documented, on par with the standards in conventional securities markets, which reflects the substantial heterogeneity in the content of these documents (Howell et al., 2020; Zetzsche et al., 2017).

Ultimately, the discretionary and largely heterogeneous nature of these documents raises the question of whether white papers contain credible information, resonating with a longstanding debate in financial accounting between voluntary versus mandatory disclosures (see, Healy and Palepu, 2001). Theoretical research argues that the disclosure of information leads to liquid and efficient financial markets, resulting in a lower cost of capital for firms (Grossman and Hart, 1980; Milgrom, 1981). A central prediction of information economics is that these incentives would lead ICO issuers to provide information relevant to investors voluntarily. However, as often observed in traditional markets, without oversight, in non-ideal conditions, firms are disinclined to make adequate disclosures (see, Beyer, Cohen, Lys, and Walther, 2010). In fact, Adhami et al. (2018) observe no significant relationship between the availability of white papers on the probability of reaching the stated funding goal. Similarly, Momtaz (2020a) observes only a marginal relationship between the number of words in white papers and the time-to-funding and finds that white paper length is unrelated to the amount raised, the time-to-listing, the first-day return, or the first-month token-price volatility. Moreover, his study shows that token issuers systematically exaggerate information disclosed in white papers, often undetected by investors, leading to more favorable outcomes for ICOs with exaggerated white papers. Furthermore, in a study by Florysiak and Schandlbauer (2021), the authors find that external factors, such as the number of industry peers, have no influence on the document's informative content, therefore, finding no meaningful link between market conditions and information in white papers.

In contrast, there is also evidence suggesting that ICOs accompanied by white papers are more likely to succeed. For instance, Bourveau et al. (2021) study 2,113 ICOs and find that lengthier and more technical white papers that disclose information about the team, token incentive structures, and governance measures (e.g., token vesting and lock-up) positively predict successful capital raising.

[4] Zetzsche et al. (2017) find that roughly one-fifth of the white papers in their sample did not contain any information about the issuing entity. Among the white papers with firm names, only 32.9% mentioned the country-of-origin and provided a postal address.

The authors conclude that white paper disclosures are relevant to the investment decision, which complements the findings of other studies showing that longer white papers tend to attract more investment and have a higher likelihood of token issuance (see e.g. Howell et al., 2020; Amsden and Schweizer, 2018; Fisch, 2019). Stylistic attributes of the text in white papers have also been shown to influence ICO outcomes (Zhang et al., 2019, 2021; Samieifar and Baur, 2020; Feng et al., 2019; Dittmar and Wu, 2019; Sapkota and Grobys, 2021). Furthermore, there is evidence in prior research that shows that ICOs with white papers that discuss the proposed technology, the technical features and benefits of the supporting blockchain architecture tend to be more successful (Howell et al., 2020; Zetzsche et al., 2017; Fisch, 2019; Barraza, 2019; Lyandres et al., 2020). Given the contrast in findings between the two groups of studies, the extent to which white papers are informative to investors is unclear and remains under-examined.

2.3. Research Questions – Unpacking the black box of white papers

Despite the ongoing debate on the severe level of asymmetric information in this market, there is an important gap in the literature. Disclosure attributes of white papers, such as tone or readability, are repeatedly highlighted as key determinants of ICO success. However, measures used to represent information content have been weak, often based only on the count of words in the text. Such disclosure metrics are arguably bad proxies of document informativeness because they overlook the context and the type of information conveyed in a document. In fact, such linguistic attributes make abstraction of the underlying meaning and consider words as independent and informative units (Loughran and McDonald, 2016). In addition, while the analysis of tone or readability could arguably be applied to analyze standardized audited financial corporate disclosures (e.g., 10-K), this approach to measuring informativeness is particularly problematic for ICO white papers given the level of heterogeneity and the lack of a formal control mechanism or standardization. The flexible nature of disclosure content requires a more extensive set of dimensions along which we could analyze white paper narratives (Hoberg and Lewis, 2017).

We propose that the information content of white papers depends not only on the quantity of information but also on the type of information provided in the white paper. While considering investing in an ICO, an investor seeks information about various aspects of the project, including information on the technicalities, the team of managers, and the projected profitability. As such, we suggest that white paper information needs to be represented on the basis of the granular concepts

that it represents and not on one-dimensional measures, such as the count of words it contains. The studies of Fisch (2019), and Bourveau et al. (2021) are arguably the closest paper to ours in attempting to understand the content of ICO white papers. While both Fisch (2019) and Bourveau et al. (2021) rely on human coders to classify the content of white papers into pre-specified items, we distinguish ourselves from these paper by conducting a holistic and replicable method, namely topic modeling, that identifies the topics covered in white papers without imposing restrictions based on what we expect to find *ex ante*. Topic modeling also allows us to highlight non-technical topics, which are overlooked in prior literature, and also to illustrate the large heterogeneity of themes covered in white papers. In addition, we set out to further leverage the information contained in the thematic content to enhance our ability to identify the topics with value-relevance to investors.

We first consider topic modeling to summarize a collection of white papers into individual topics (representing concepts). Topic modeling allows for a richer analysis than the previously popular approaches of representing information through keywords (Zhang, Jin, and Zhou, 2010). We then test the impact of topics on the overall influence of a white paper on the ICO outcome. We expect the topics contained in white papers to improve the identification of successful ICOs, beyond what can be achieved using quantitative financial metrics and aggregate measures of textual style features. We, therefore, contribute to the literature by addressing the following research questions:

Research Question 1: What are the themes discussed in ICO white papers?

Research Question 2: Is the thematic content of white papers informative to investors and consequential to ICOs' outcome, relative to other ICO-specific characteristics and white paper texts' style features?

Research Question 3: Which topics contained in ICO white papers are associated with successful ICOs?

ICO fundraising is the first milestone for a successful blockchain-based project. A following key question is whether the white paper's informational content is indicative of the project's future performance. A reason to expect a lasting influence of white papers is that, given its primary purpose to persuade investors, these documents may contain crucial forward-looking information, illustrating future profitability and earnings for investors (Dittmar and Wu, 2019). Such information detailing project road-map and targeted milestones set expectations among investors and are likely to maintain

the ICO white paper's relevance beyond the ICO's completion. Moreover, certain topic communications, for instance, those that reflect the project's technology or target market, are a correlate of an underlying project quality, which may have a persisting relationship with the project's future performance. In contrast, white papers' influence could also substantially diminish after the token is listed as supply and demand dynamics start to form equilibrium prices based on all the available information about the firm (Momtaz, 2020a). The token's exchange listing after a successful ICO performs the coordination function that dispersed investors fail to perform on their own during the ICO (see, Kremer, Mansour, and Perry, 2014). The token price on exchange platforms reflects the crowd's aggregated wisdom, which becomes available to all dispersed investors through the readily-observable market price, leading to an aggregated pricing signal. Therefore, relative to the ICO stage, the tokens' listing reduces the level of information asymmetry between investors and managers, diminishing the information value of white papers in the post-ICO period. Given this backdrop, which thematic components of white papers remain relevant, while which ones are no longer consequential to the project's post-ICO performance constitute empirical questions of interest. We formulate our research inquiry as follows:

Research Question 4: Is the thematic content of white papers informative of the project's post-ICO performance, relative to other ICO-specific characteristics and white paper texts' style features?

Research Question 5: Which topics contained in ICO white papers explain post-ICO performance?

3. Method and variable definitions

This section describes the method and variables used in our empirical tests. We first elaborate on the sentLDA method used to map the thematic content of ICO white papers. We then detail the model, and test the identified thematic content's information value in explaining ICO outcome and post-ICO performance.

3.1. Implementing sentLDA for knowledge extraction from ICO white papers

Implementing a machine learning technique for unsupervised knowledge extraction provides notable advantages compared to traditional methods involving human coders. In contrast to the labor-

intensive manual approach, an automated data-driven methodology is not subject to idiosyncratic biases, and offers replicability along with exponentially higher computational capacity (Morris, 1994). One such automated tool that researchers widely use in the information retrieval literature is Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). Using multiple documents, LDA infers the distribution of topics, each of which is represented by a distribution of words. Employing human validation techniques, multiple studies show that the computed thematic results from LDA are semantically meaningful and correspond to human interpretations (Chang, Gerrish, Wang, Boyd-Graber, and Blei, 2009; Huang, Lehavy, Zang, and Zheng, 2018; Dyer, Lang, and Stice-Lawrence, 2017).

Topic modeling tools, and in particular LDA-related techniques, have been successfully applied to textual documents to operationalize various latent firm attributes that otherwise pose identification and measurement challenges. For instance, in a recent study, Bellstam, Bhagat, and Cookson (2020) develop a measure of innovation using the text in analyst reports of S&P 500 firms without relying on patenting and R&D information, allowing the researchers to identify innovative non-patenting firms (see also, Kaplan and Vakili, 2015). Similarly, Brown and Hillegeist (2007) and Hoberg and Lewis (2017) employ the technique to respectively identify misreporting and abnormalities in the communication by fraudulent managers. Furthermore, topic modeling has also been used to examine analyst reports and conference calls (Huang et al., 2018; Giorgi and Weber, 2015), stock market movements (Curme, Preis, Stanley, and Moat, 2014) and corporate risk disclosures (Campbell, Chen, Dhaliwal, Lu, and Steele, 2014; Bao and Datta, 2014) (for an overview, see Eickhoff and Neuss, 2017).

To identify and quantify the thematic content in white papers, we rely on a specific iteration of LDA, namely Sentence Latent Dirichlet Allocation (sentLDA), initially proposed by Bao and Datta (2014). We employ this method because it relaxes the bag-of-words assumption in traditional LDA and imposes an additional constraint that all words in a sentence derive from a single topic. The traditional LDA allows the user to obtain the relative occurrence of topics in a document but does not specifically indicate where, in which section or sentence, the individual topic is located. Including this sentence boundary, however, allows us to obtain a more fine-graded level of topic allocation within each document and increases our understanding of which sentences pertain to a specific topic. For our study, sentence structure information is especially relevant, as we are concerned with identifying the various topics within white papers shared by ICOs of all types as opposed to classifying the documents (see Fu et al., 2019).

Much like the traditional LDA, the sentLDA model relies on a few simple assumptions. It assumes a

collection of K topics in a given document d and that the list of words in each topic follows a Dirichlet distribution, $\beta_k \sim \text{Dirichlet}(\eta)$. Furthermore, for each document, sentLDA considers a vector of topic proportions drawn from a Dirichlet distribution $\theta_d \sim \text{Dirichlet}(\alpha)$. Relying on these assumptions, in addition to a few learning parameters, the sentLDA model categorizes words in each document into K number of topics and assigns a specific topic to each sentence in a document. The critical inferential problem here is computing the posterior distributions of the two hidden variables θ (the topic proportions for each document) and z (the topic assigned to each sentence), given the model parameters and the observed documents. As the distribution is intractable (Blei et al., 2003), we need learning algorithms to approximate the posterior distributions. Accordingly, we follow Bao and Datta (2014) and use the Variational Expectation Maximization learning algorithm. In this fashion, sentLDA organizes the words in a collection of documents into a dictionary of topics and defines each document as a collection of sentences, each belonging to a specific topic (for further details on the computation, see Bao and Datta, 2014).

3.2. Regression analyses examining the predictive value of sentLDA white paper topics

To investigate the information value of the thematic composition of white papers, we use the K topics obtained from the sentLDA model as input variables in a multi-factor model. Specifically, we examine the relationship between the white papers' topic content with ICOs' success and the issuing project's post-ICO performance.

The benchmark model is the traditional approach consisting of a linear model in which the ICO outcome and post-ICO performance are estimated using white paper-, ICO- and country-specific variables:

$$Y_j = \alpha + \beta \cdot WhitepaperControls_j + \beta \cdot CountryControls_j + \gamma \cdot ICOControls_j + \varepsilon_j, \qquad (1)$$

where ε_j are country-clustered standard errors, corrected for heteroskedasticity. Y_j represents the several ICO outcome variables, as well as the post-ICO performance variables. As measures of ICO outcome, we consider whether the issued token was listed on a secondary exchange, the amount raised and the time-to-listing. The post-ICO performance variables are the abnormal first-day initial returns, 30-day realized volatility, 6-month buy-and-hold abnormal returns, and if the issued tokens were delisted. We detail each of the dependent variables below (see Section 3.2.1 and 3.2.2). To

measure the information value of the thematic content of ICO white papers, we include the topics estimated based on the sentLDA ($Topic_{k,j}$) and define the following model:

$$Y_j = \alpha + \delta_k \cdot \sum_{k=1}^{K} Topic_{k,j} + \beta \cdot WhitepaperControls_j + \beta \cdot CountryControls_j + \gamma \cdot ICOControls_j + \varepsilon_j, \tag{2}$$

where ε_j are country-clustered errors, corrected for heteroskedasticity. $Topic_{k,j}$ is defined as the number of sentences referring to topic k in the white paper of ICO j.

To test the incremental information value of white papers' thematic content, we compare the two regressions based on the models' out-of-sample predictions using bootstrap resampling with 1,000 replications. Bootstrap resampling is a common approach for evaluating the out-of-sample performance of regression models, and is used to improve model stability and avoid overfitting (Efron and Tibshirani, 1994; Hastie, Tibshirani, and Friedman, 2009). It involves generating multiple training sets based on uniform re-sampling with replacement of rows in the dataset. On average, the random re-sampling with replacement results in 63.2% of the original sample, and each bootstrapped sample serves as the training dataset for the respective iteration. The remaining observations are used as the out-of-sample test set. As such, the model is trained and its performance is estimated on the out-of-sample sets (Tsamardinos, Greasidou, and Borboudakis, 2018).

3.2.1. ICO outcome variables

Token Traded. Following Amsden and Schweizer (2018) and Howell et al. (2020), we use a binary variable indicating whether the issued tokens are traded in a secondary market. Popular measures of success used in other modes of financing, such as successfully raising the goal amount (as in crowdfunding), are not feasible in the context of ICOs, as issuers are not compelled to specify a funding target. In fact, we find only 74.37% of the ICOs specify a soft-cap, the term used for pre-set funding target in ICOs.[5] However, all ICOs look to issue tradeable tokens, irrespective of the nature of business or ICO-specifications. Hence, we consider an ICO to have successfully culminated if the issued tokens are listed on Coinmarketcap.com because such a listing acts as a notable industry

[5]Soft-caps can be viewed as equivalents to goal amounts in all-or-nothing crowdfunding. If the soft-cap is not met by the end of the ICO, contributors are automatically reimbursed.

validation for newly founded projects and tokens.[6] We specifically consider Coinmarketcap because it is the leading price-tracking website, which applies stringent listing criteria.[7] We employ a logistic regression to estimate the model with the binary success measure.[8]

Amount: In order to distinguish the magnitude of success, we incorporate models with the amount raised as the dependent variable (see e.g., Fisch, 2019; Adhami et al., 2018). While the variable does not indicate if the project managed to meet its specific fundraising goals, it is particularly relevant for our analyses, given that it provides a reasonable indication of the project's popularity among investors. However, a particular draw-back using this variable is that, since ICOs are not compelled to make such information public, the information is not available for a substantial number of ICOs. Consequently, our respective analyses are restricted to a smaller sample. Furthermore, to mitigate the influence of outlying observations raising exceptionally high amounts, we take the natural logarithmic values of the *Amount* variable.

Time-To-Listing: As our third measure of ICO success, we look at the duration between the end of the ICO and the issued tokens' listing. The first milestone for a project after a successful ICO is to list the tokens to investors, thereby providing them with an exit option and increasing the token's liquidity (Lyandres et al., 2020). Therefore, the period of time to become listed indicates how the project fared during the ICO process. Projects that meet fundraising targets are likely to issue the tokens sooner. Similar to Momtaz (2020a), we define the variable as the number of days between the start of the ICO and the date its token is listed on Coinmarketcap. For the remaining ICOs without listed tokens, we take the number of days from the ICO start date to July 20, 2021–the date on which the price data were collected. As the variable has a time-to-event structure, we employ a Cox Proportional-Hazards model.

3.2.2. Post-ICO performance indicators

Because issuing projects are not obligated to make financial disclosures concerning their earnings and assets, direct measures of the projects' post-ICO performance are not available. Instead, we look at

[6] To operationalize the *TokenTraded* variable, we rely on historical data provided by Coinmarketcap, which allows us also to identify tokens that were listed but later removed by the platform.

[7] For details, see https://support.coinmarketcap.com/hc/en-us/articles/360043659351-Listings-Criteria

[8] It should be noted that although the model is restricted to ICOs that have a white paper, we do not apply a selection model since there are several distinct factors determining the availability of white papers. While some ICOs may not have had a white paper during the ICO, many other projects remove their white papers from their websites after the completion of the ICO. Furthermore, some white papers are in languages other than English, and we also find some white papers in image formats that are not computer-readable. Given the lack of a consistent factor that adequately describes the selection process and also satisfies the exclusion restriction criteria, for our main analyses, we resort to sub-sample analyses as recommended by Puhani (2000).

token-price and listing indicators as proxies for project performance as practiced in the past literature (e.g., Howell et al., 2020; Momtaz, 2020a; Fisch and Momtaz, 2020).

Abnormal Initial Return: A number of studies examine investors' initial reaction to the token as it is traded for the first time (e.g., Momtaz, 2020b; Felix and von Eije, 2019; Benedetti and Kostovetsky, 2021). Apart from indicating the degree of underpricing, which is documented as substantial among ICOs, the first-day-return also reflects the immediate market response to the newly issued token. Furthermore, as decentralized projects count on generating early network effects, the initial price momentum could be crucial for the project's long-term success (Momtaz, 2020a). We therefore examine if the thematic composition of white papers is associated with initial returns from the issued tokens. Given the evidence of notable collinearity in the cryptocurrency market (Katsiampa, 2019; Qiao, Zhu, and Hau, 2020), we follow Momtaz (2020a), and take the market adjusted value of initial returns, formulated as follows:

$$AbnormalIR_i = \frac{P_{i,t=1} - P_{i,t=0}}{P_{i,t=0}} - \sum_{j=1,j\neq i}^{n} \frac{MC_{j,1}}{\sum_{j=1,j\neq i}^{n} MC_{j,1}} . \frac{P_{j,t=1} - P_{j,0}}{P_{j,0}}, \tag{3}$$

where the market capitalization-weighted market return is subtracted from the individual token price return on the first day.

Initial Volatility: The volatility in the token price returns observed in the first month of trading represents the investors' uncertainty about the project during the period. From the existing literature, we know that some ICO attributes, such as having a pre-ICO, the extent of exaggeration in white paper text, contribute to token-price volatility (Howell et al., 2020; Momtaz, 2020a; Lyandres et al., 2020). We further examine which specific white paper topic components associate with the realized volatility in token prices during the first 30 days from the listing. We compute initial realized volatility as follows (Andersen, Bollerslev, Diebold, and Labys, 2003):

$$InitialVolatility_i = \sqrt{\sum_{t=1}^{30} [log(\frac{P_{i,t}}{P_{i,t-1}})]^2}. \tag{4}$$

Abnormal Long-term Returns: Concerning the issuing project's long-term performance, we consider

buy-and-hold abnormal returns for investors holding the token for 180 days after the first day of trading (for examples of similar application, see Benedetti and Kostovetsky, 2021; Lyandres et al., 2020; Fisch and Momtaz, 2020). Since ICOs remain a novel phenomenon and most listed tokens do not possess price data for substantial durations, we examine the projects' long-term performance using realized returns during the first 6 months of trading (in line with Lyandres et al. (2020) and Fisch and Momtaz (2020)) Furthermore, similar to our measure of market adjusted initial returns, we follow Momtaz (2020a) and Momtaz (2020c), and correct for market capitalization-weighted market returns, which account for the influence of market fluctuations on individual token returns. We formulate the measure as follows:

$$AbnormalLR_i = \frac{P_{i,t=180} - P_{i,t=0}}{P_{i,t=0}} - \sum_{j=1,j\neq i}^{n} \frac{MC_{j,t=180}}{\sum_{j=1,j\neq i}^{n} MC_{j,t=180}} . \frac{P_{j,t=180} - P_{j,t=0}}{P_{j,t=0}}. \tag{5}$$

Delisted: Delisting is defined as an event when a once-listed token is no longer listed on Coinmarketcap, which we consider as a proxy for post-ICO project failure (see also Momtaz, 2020a). There are several reasons as to why coins get delisted, including low liquidity, cessation of business activity, poor project implementation, and legal charges.[9] In the absence of organized public information on the status of the issuing firm, the variable provides a reasonable indication of whether the ICO project failed after having successfully issued and listed its tokens.

Note that, since we derive the four post-ICO performance indicators from the token's price or listing status, our empirical analyses of these models are limited to projects that successfully issued their tokens and got listed. Given that a selection criterion leads to these test samples, we implement Heckman's two-stage correction (Heckman, 1979) to estimate the post-ICO performance models. The first stage involves estimating a probit model for the likelihood that the project's token is listed, equivalent to our first model for ICO success. In the second stage equation, we include the inverse Mills ratio calculated from the density and distribution functions from the first stage to sample the selection. As Heckman selection models require exclusion restrictions that explain the selection process but do not directly influence the outcome (Bushway, Johnson, and Slocum, 2007), we omit the control variables specifically related to the ICO's outcome from the second stage.

[9]For a detailed description of Coinmarketcap's delisting criteria, see https://support.coinmarketcap.com/hc/en-us/articles/360043659351-Listings-Criteria

3.2.3. Control variables

To obtain unbiased estimates of the effect of white paper topics on the ICO's success, we include a number of control variables. Given the novelty of the literature, the list of control variables included largely varies across studies, and a consensus in this regard has yet to emerge. Nonetheless, we draw inspiration from the works of Fisch (2019), Adhami et al. (2018) and Amsden and Schweizer (2018) in selecting a list of prominent ICO-level and country-level control variables.

We include 20 control variables in addition to year-quarter time dummies and region dummies to account for various attributes relating to the white paper text, the project and the ICO that could potentially influence our outcome variables. The $WP\_Readability$ variable represents the readability of the white paper text based on the Gunning-Fog Index. The measure, which is derived from a linear combination of average words per sentence and the proportion of complex words (words with more than two syllables), quantifies the difficulty in reading a document. We also control for the tone in the white paper text using the bag-of-words approach based on the Loughran and McDonald (2011) dictionaries for positive and negative tone ($WP\_Sentiment$). The variable represents the proportion of net positive words (positive minus negative words) in the white paper. As a measure of total length of the document, we control for the number of pages in the white paper PDF document (WP_Pages). Furthermore, to account for the variety of information in white papers, we further control for the overall white paper topic diversity based on the Shannon diversity index ($TopicDiversity$).[10] Apart from the document attributes, we also control for the relevant country-specific attributes. First, we include a dummy variable indicating if the project is based in a country considered to be a tax haven (Hines, 2010). Second, following Shrestha et al. (2021), we include an aggregated measure for institutional quality in the project's host country based on Worldwide Governance Indicators (Kaufmann, Kraay, and Mastruzzi, 2010).

Apart from the white paper and country-level characteristics, we control for various project attributes, which include dummy variables indicating if the ICO restricts the participation of investors based in the US ($USRestrict$), if the issued tokens are built on the Ethereum platform (Eth), whether the project implements the Whitelist or Know-Your-Customer (KYC) guidelines ($WhitelistKYC$), if the issuer details the proportion of the total tokens to be distributed ($TokenDist$), and if the project

[10] Shannon Diversity Index is defined as $-\sum_{j=1}^{30} P_{j,i}.lnP_{j,i}$, where, for campaign i, P_j is the percentage of sentences in a white paper dedicated to topic j out of 30 topics (if $P_j = 0$, lnP_j is set to 0) (Shannon and Weaver, 1963). The index quantifies the uncertainty in predicting the identity of a randomly chosen entity from a given data (for examples of application, see Reguera-Alvarado, de Fuentes, and Laffarga, 2017; Campbell and Minguez-Vera, 2008).

provides a link to a presentation video ($Video$). Additionally, the models include a variable representing the standardized average rating the project receives from the various sample sources ($Ratings$). We also control for the number of team members the project lists in the white paper ($Team$). As a proxy for the project's social and promotional reach, we include a variable indicating the number of social media pages the project has ($SocialMedia$).

Moreover, we include a number of ICO-specific attributes that are regularly used in the empirical literature. Specifically, our models include a dummy variable indicating if the ICO specified a minimum investment requirement ($MinInvest$), if contributions in fiat currencies were accepted ($Fiat$), whether a pre-ICO was organized before the ICO ($PreICO$), if a hard cap or soft cap was specified ($HardCap$ and $SoftCap$), and if the issuers offered bonus schemes ($Bonus$). Furthermore, we also include a variable indicating the number of currency options that the ICO provided to investors ($NumbCurr$). As specified, we omit the last seven ICO-specific variables from the second stage of the two-stage models concerning post-ICO performance. We briefly define all variables in Table 1.

< Insert Table 1 about here >

4. Data and summary statistics

4.1. Data

Our sample is compiled from seven prominent ICO-listing websites, namely ICOBench.com, ICO-Holder.com, ICOMarks.com, ICORatings.com, ICODrops.com, FoundICO.com and CryptoCompare.com. Gathering ICO data for empirical analyses is particularly challenging as ICOs allow firms to circumvent intermediaries. Projects can directly provide relevant information on their websites alone, and therefore, a centralized repository with details of all ICOs does not exist. However, given the growing interest in ICO investments, third-party ICO-tracking websites have emerged, which offer detailed information on considerably large pools of ICOs. The list of token offerings in these websites, however, are not exhaustive, and other issues concerning potential errors, absence of unique identifiers, and the lack of consistent and updated information are known drawbacks (Lyandres et al., 2020). Therefore, we rely on multiple ICO tracking websites, allowing us to obtain a more complete overview of the ICO universe. This approach substantially diverges from prior literature, which often relies on a single ICO data source (e.g. ICOBench) (see Momtaz, 2020a; Fisch, 2019; Amsden and

Schweizer, 2018; Howell et al., 2020). Benefiting from multiple sources, we also cross-verify details specific to each ICO and identify the most reliable information when there are inconsistencies between sources. We detail the data compilation process in in Section A1 of the online appendix.

From seven ICO tracking websites, we gather a dataset consisting of 9,159 unique ICOs launched between July 2012 and July 2021. From these ICOs, we find computer-readable white papers for 5,897 ICOs.[11] Since our study centers on the semantic content in these documents, we then drop the documents that are not in English and those that are in unsuitable formats (such as, picture format and without sentence separators). We obtain a sample of 5,210 observations.

For the first step of our analysis, i.e., the application of sentLDA to estimate the types of topics discussed, we use this set of white papers. Here, we proceed with several common filtering steps to optimize the topic modeling procedure. Using a standard set of English language stop-words, we remove highly frequent words, such as 'is,' 'the' and 'and,' which have little standalone thematic meaning. Then, based on tf-idf (term frequency-inverse document frequency) scores, we filter out words that are specific to a given document, since our focus is to identify common themes across white papers. Furthermore, to ensure that we exclude erroneous spellings, numbers, and terms specific to a document, we remove all words not in the dictionary of English words (UK and US) and in the list of 200 most frequent words in the entire collection (which captures novel terms, such as blockchain and Ethereum, that may not be in the dictionaries). The set of white papers consists of an aggregate vocabulary of 35,295 words.

In the second step of our study, which involves regression analyses on the relationship between white paper topics and our dependent variables, we implement additional filters. While the set of white papers used for sentLDA contains useful information to train the model, for our regressions, we remove outlying and non-relevant observations, and those that have incomplete information. First, we exclude observations that are not characterized as ICOs but as other emerging token offerings types, such as Initial Exchange Offerings (IEOs) and Security Token Offerings (STOs), which have distinct functional forms.[12] Second, to ensure that the white papers in our sample reflect the information that was available to investors during the token offering, we remove white papers that were modified after the ICO end date based on the last modification date available in the white paper PDF files'

[11] It should be noted that although we do not find all white papers, it does not *ipso facto* imply that these projects did not have a white paper during their ICOs. Since we collected the white papers after the ICO, we found that several ICOs no longer had a working link to their white papers, which may be the result of changes in the project's website.

[12] We include these white papers in the topic modeling sample, as these white papers can still inform the topic identification process. Nevertheless, we drop these observations from the regression analyses, as the decision criteria of investors can vary across the different financing settings, given the differences in third-party oversight and investor incentives (Miglo, 2020, 2021).

metadata.[13] We then drop extremely short white papers (single-paged, 30 sentences, and/or less than 3,000 characters), which ensures we exclude documents in partial-image formats. We also drop all observations with missing control variables, and we trim the data at the 1^{st} and the 99^{th} percentile based on the continuous dependent variables to reduce the influence of extreme observations. Our final study sample consists of 2,505 ICOs ending before June 1, 2021. The data compilation steps are detailed in Table A7 of the online appendix.

4.2. Summary statistics

We report the summary statistics of the dependent and control variables in Table 2. We find that 18.64% of the projects in our sample issued tokens that were eventually listed on Coinmarketcap. ICOs raise, on average, an amount of $8.85 million. Among the tokens listed, we find that it typically takes seven months to get listed after the end of the ICO. On average, the abnormal first-day return is 13.98%, the realized volatility observed in the first trading month is 84.37%, and the abnormal returns during the first six months equals −40.30%. Moreover, we find that 34.98% of our sample tokens are already delisted from Coinmarketcap.

< Insert Table 2 about here >

Among the control variables, we observe that the average readability score is 15.59, which lies within the range of difficult to read text that is characterized as suited for college graduates and is comparable to the scores observed in Bourveau et al. (2021) and Samieifar and Baur (2020). The average positive sentiment is 0.28%, ranging between -2.53% and 3.03%, and the average document contains about 33 pages. We find that the average diversity score is 2.23, which indicates that, generally, the documents are fairly diverse.[14] 33.89% of projects are from countries that are recognized as tax havens. From the institution scores, we know that ICOs have been launched in countries scoring both high and low in terms of institutional strength; however, the mean institutional score indicates that most ICOs are based in countries with strong institutional quality. Furthermore, we find that 39.36% of the projects imposed restrictions on investors based in the US, a substantial majority of the projects are based on the Ethereum blockchain (85.51%), 65.95% of the ICOs implement either Whitelist or

[13] White paper versions that were modified after the ICO's culmination mainly appear due to changes by issuers to inform the stakeholders of new updates.

[14] The maximum diversity score possible is ln(k). Since k = 30 in our case, the maximum value is 3.40, which occurs when all topics are equally distributed in a risk disclosure.

KYC protocols, and 94.41% provided details concerning the token distribution scheme. Moreover, the average number of team members in a project is approximately 12.41. Concerning the ICO-specific characteristics, we observe that 43.43% of the ICOs specified a minimum investment requirement, more than two-thirds launched a pre-ICO (68.66%), 85.59% specified a hard cap, 74.37% specified a soft cap, 64.75% included a bonus scheme. A typical ICO offers around two currency options for investors, and only 24.19% ICOs included fiat payment as an option. Panel B of Table 2 presents the correlation matrix for all our dependent and control variables. Several variables are significantly correlated, and thus we employ multivariate analyses for all our tests.[15]

5. The topic modeling of ICO white papers

This section provides the results addressing our research questions. We proceed in three steps. Before analyzing the results, we first discuss the evaluation of the topic modeling approach. We then provide details on the topics contained in the ICO white papers. Finally, we test whether the topical content of ICO white papers serves as a determinant of ICO success and post-ICO performance.

5.1. Implementation of topic modeling

We run the sentLDA algorithm on the corpus of cleaned white paper documents to generate a list of 30 topics. As a probabilistic model, sentLDA assigns weights corresponding to each topic to every word in the vocabulary. Thus, the topics are defined as sequential lists of words based on the topic-weights assigned. sentLDA then allocates each sentence in a white paper to the highest weighted topic based on the words it contains. The output can be described as follows:

$$Topic_k = TopicWeight_k.Word_z, \tag{6}$$

$$TopicAssignment_S = k|\max_k \sum_{w=1}^{W} TopicWeight_k.Word_w, \tag{7}$$

where k represents the k^{th} topic, z is a word in the total vocabulary, s is a given sentence in a white paper, W represents the total number of words in sentence S, and w represents the w^{th} word

[15]Table B5 of the online appendix further provides the correlations between the various topic categories and other text-based variables in our models.

in sentence S. Therefore, for every ICO, sentLDA provides a vector output of 30 elements, which describes the distribution of the per-sentence topic allocations.[16]

Because sentLDA is unsupervised, it is necessary to first evaluate the algorithm's effectiveness in capturing human comprehension. To do this, we follow prior studies and combine human and automated evaluation methods on the meaning and interpretability of the topics inferred from the ICO white papers' narratives. Given the manual process of topic labeling, one limitation of this approach is that it naturally involves human judgment. This limitation, nonetheless, is not a particular concern as the literature on the evaluation of unsupervised topic modeling methods, particularly LDA, emphasizes the semantic interpretability of the results over statistical measures (Grimmer and Stewart, 2013; Chang et al., 2009; Bao and Datta, 2014; Bellstam et al., 2020; Hoberg and Lewis, 2017; Huang et al., 2018).

To determine a meaningful interpretation for each topic, we first generate a list of the highest weighted phrases and sentences for each topic. Specifically, we construct lists of 1,000 sentences per topic based on the weights assigned to their constituting words. Next, we sort the sentences by length and extract the middle tercile (334 sentences) as representative sentences of typical length. We also extract the 20 most frequent bigrams (two-word phrases excluding stop words, numbers and symbols) from the 334 mid-length sentences. These sentences are also sorted based on the cosine similarity between them. We then evaluate the semantic meaning of the top 20 bigrams and the top 100 mid-length sentences based on cosine similarity and assign descriptive labels to each topic. Table 3 provides the word clouds representing the list and weights of the highest weighted words in each topic with their associated labels, while Table 4 provides the labeling of each topic and briefly describes the 30 topics.[17]

< Insert Tables 3 and 4 about here >

As a second evaluation method to validate our classification, we graphically examine the similarities between the topics based on the words they constitute. Figure 1a displays a network graph where each weighted line represents the correlation between adjoining topics based on the weights assigned

[16] A point to note is that although sentLDA is an unsupervised method, we are required to specify the number of topics ex-ante, similar to the traditional LDA. We select 30 topics based on comparative analysis with outputs from 15, 20, 25, 30, 35, 40 and 50 topics. We find that the resulting topic-words display superior cluster quality and greater semantic coherence (i.e., it reveals the maximum number of distinct themes while minimizing the overlaps between topics).

[17] Tables B2, B3 and B4 of the online appendix provide the lists of 20 highest weighted words, the 20 most common bigrams, and the 100 mid-length representative sentences for each topic. This form of evaluation is qualitative and is supported by several studies, including those by Hoberg and Lewis (2017), Dyer et al. (2017) and Brown and Hillegeist (2007).

to all the words (for ease of interpretation, correlations under 35% are not displayed). It shows the degree of overlap between topics based on the words they emphasize. We observe several clusters. For instance, topics *HumanResource* and *Expertise* display a notable degree of correlation (40.05%), meaning that the sentences discussing the project's team and their job description often include words relating to illustrations of specific skill sets. Likewise, we find clusters between topics concerning risks and regulations, such as *LegalDisclaimers*, *RiskManagement* and *RiskDisclosure*, and also between topics related to various blockchain-specific technical details, such as *BlockhainEncryrption*, *Governance* and *SmartContract*. Since our topic labels are discretionary, these linkages provide additional nuance and support for the interpretation of the topics.

< Insert Figures 1a and 1b about here >

Together, our evaluation methods suggest that the sentLDA algorithm provides a valid set of semantically meaningful topics that are reasonably coherent and interpretable by human judges. The following section and the corresponding table describe the salient features of the estimated thematic composition obtained from our sample white papers; thus, addressing *RQ1*.

5.2. The thematic content of ICO white papers

Table 4 also reports the summary statistics of the various topic variables, and the 10 grouped topic categories. For ease of interpretation, we manually group the 30 topics into 10 categories, representing distinct themes that capture the common thread across their constituting topics. In devising these categories, we first rely on the clustering observed in Figure 1a and then allocate suitable labels to represent the broader themes. We refer to these aggregate topics as "topic categories". Among the 10 categories, we find that *ICO*, *Product*, *Profitability* and *Network* are the most prominent; in contrast, *People*, *Innovation* and *Risk* are the least frequent. We find that *BlockchainApplication* (16.80 sentences) is the most frequent topic in ICO white papers, while *RiskDisclosure* (2.45 sentences) is the least observed. Other topics such as *Expertise*, *HumanResource*, *ServiceProfile* and *FinancialServices* are distributed around the sample average of 10 sentences. The standard deviations of most topics are substantial in comparison to the mean, indicating much variability in the information composition of the white papers. As none of the topics appear in all white papers, and given the high observed maxima of some topics, we can also deduce that some documents are dominated by a specific topic. Overall, we find that an average document contains about 19 unique

topics.

We further examine which topics are more likely to appear in the same white paper. Figure 1b reports a network graph with each weighted line indicating the degree of co-occurrence between topics within documents (for ease of interpretation, correlations under 15% are not displayed). We observe, for example, that the topic $HumanResources$ regularly appears with sentences illustrating team members' expertise ($Expertise$) (Correlation: 45.32%). Also, sentences discussing technical specifications of the underlying blockchain, such as smart contract application ($SmartContract$), block validation mechanism ($ConsensusMechanism$), encryption protocols ($BlockhainEncryption$) and data management ($Data\&AI$) are often found in the same white papers. There are also visible clusters of topics concerning risk factors ($RiskDisclosure$, $RiskManagement$, $Terms\&Conditions$ and $LegalDisclaimers$), and topics concerning ICO details ($InitialSale$, $TokenBenefits$ and $BuyIn$). Among all topics, the two that appear the most frequently together are $Regulations$ and $RiskManagement$ (Correlation: 67.95%). Furthermore, the size of each node is calibrated to reflect the total number of topic sentences in the overall sample. The topics $BlockchainApplication$, $PlatformDevelopment$ and $Data\&AI$ are the most prevalent (81,287; 75,228; and 64,952 sentences, respectively); whereas, the topics $RiskDisclosure$, $LegalDisclaimers$ and $Terms\&Conditions$ (11,309; 21,919; and 25,451 sentences, respectively) appear the least.

Furthermore, in Figure 2, we illustrate how white paper topic content has evolved over time, which shows the topics that are consistently featured in white papers, and also helps us identify time-specific trends on which types of information are discussed in ICO white papers. It displays the average number of topic sentences in each quarter from April 2017 to December 2020.[18] We find that while the topic categories ICO and $Profitability$ show consistently high presence, the discussions on $Innovation$ is relatively low, especially in more recent periods. We also find that the discussions on topic categories $Mining$ and $Product$ are decreasing in more recent white papers, whereas, $Risk$ topics are more prominent. Furthermore, we observe that the presence of topic categories $Blockchain$, $Security$ and $People$ is rather consistent throughout the sample period.

< Insert Figure 2 about here >

[18] We do not plot the periods before and after the specified duration, as they each contain less than 10 observations.

5.3. The information value of ICO thematic content and ICO outcome

Based on the categories identified above, we now examine $RQ2$ and investigate the informativeness of the thematic content of ICO white papers by applying the multivariate model defined in Equation 2. We first report the test statistics comparing each model's out-of-sample performance to the base model in Panel A of Table 5. The models' performance measures are derived from 1,000 bootstrapped replications (Efron and Tibshirani, 1994; Hastie et al., 2009). We follow Janes, Longton, and Pepe (2009) and evaluate the difference in the average test statistics based on non-parametric Wald tests. Given that our analysis includes linear, logistic and hazard models, we follow prior literature and examine, respectively, R-squared, AUC and concordance index (Hanley and McNeil, 1982; Harrell Jr, Lee, and Mark, 1996).[19,20]

< Insert Table 5 about here >

Panel A of Table 5 provides the resulting average values of test statistics from the 1,000 bootstrap samples. Our results show that the topic variables hold incremental information value in explaining ICOs' success. For Model 1 with $TokenTraded$ as the dependent variable, we find that the out-of-sample estimated AUC increases by 0.012 (1.67%). This increase is significant at a 99% confidence level. For Model 2, which examines the relationship with the amount raised during the ICO, we find a significant increment in R-squared by 0.004 (4.14%). Model 3 focuses on the time-to-listing variable. Similar to the other variables, we observe a significant increase in the concordance measure by 0.098 (2.22%). These results collectively answer our $RQ2$ and suggest that content-based information drawn from ICO white papers improves the detection of successful ICOs beyond what can be achieved by the ICO-specific and textual-based metrics identified in prior literature.

In Models 1 – 3 of Panel B of Table 5, we report our main regression results pertaining to $RQ3$. First, among the control variables, we find that the aggregate rating score and the size of the team are significant and positively related to all the measures of ICO success. In addition, we find that the variable $Video$ has significant positive relationship with the variables $TokenTraded$ and $TimeToListing$. In contrast, we find that greater number of social media links is inversely related to the same outcome variables, suggesting that effective social media engagement is likely to be limited to fewer channels.

[19] AUC is the area under the receiver operating characteristics (ROC) curve, a standard technique for visualizing and selecting classifiers, which combines the model's true positive and false positive rates in one graph.

[20] Concordence index, or Harrell's C-index, is a widely adopted measure for assessing prediction performance in survival/hazard analysis settings (Pencina and D'Agostino, 2004). It indicates the fraction of concordant pairs in the data, i.e., the proportion of pairs where the observation with higher observed survival time also has a higher predicted survival duration (lower risk score).

Moreover, while the results show that ICOs implementing Whitelist and KYC compliance are likely to raise more funds, we also find that these projects take more time to issue their tokens. Other variables, namely *Institutions* and *WP_Readability* show significant associations with *Amount*, and *WP_Sentiment* is significant and negatively related to *TokenTraded*.

We now discuss our main findings concerning our variables of interest, the ICO white papers' topics. As reported in Models 1, 2 and 3 of Table 5, we find that not all categories share the same degree of relevance with respect to the outcome variables. Given the significant correlations, investors appear to deem some types of information as important and credible, while others seem superfluous with no meaningful relationship with any three indicators of ICO success. In particular, we find that the two topic categories that relate to the technical features of ICOs (*Blockchain* and *Mining*) are consistently and positively linked with the ICO outcome throughout all three ICO success models. This result complement the findings of Fisch (2019) and Bourveau et al. (2021), who show that the technicality of white papers tends to signal successful ICOs. Our models also highlight other topics that are consequential, but have negative relationships with the outcome variables. For instance, we find that the extent of information on *Network*, i.e., discussions on community development and promotional activities, and other topic categories, *Product* and *Risk*, have rather consistent adverse relationships. These results suggest that white papers with such communications, that are business operations-related and non-technical in nature, are often associated with lower quality projects, or may be indicating patterns and attributes that are less favored by investors.

A key highlight from our findings is that topics with significant favorable relationship with ICO success are mainly technical, which require specialized expertise and are harder to replicate. To further confirm this result and examine that the conclusions we draw are unaffected by clustering choices, we check if our results hold if we divide the topics into two clusters, i.e., *Technical* and *NonTechnical*, as shown in Panel C of Table 5. The *Technical* cluster is obtained by aggregating the sentence count of topics that concern blockchain and other technology-related topics. Specifically, the *Technical* cluster comprises of topics *InitialSale*, *BlockchainApplication*, *ConsensusMechanism*,*Governance*, *EnergySustainability*, *BlockchainEncryption*, *SmartContract*, *Data&AI* and *R&D*. We group the remaining topics into the *NonTechnical* cluster. In support of our earlier results, as shown in Panel C, we find that aggregated technical topics are particularly informative to explain the ICO's outcome, while a greater emphasis on non-technical topics is associated with a lower likelihood of the token being traded.

5.4. The information value of ICO thematic content and post-ICO performance

In Models 4 – 7 of Table 5, in Panels A and B, we provide results for $RQ4$ and $RQ5$ on the relationship between white paper topics and the issuer's post-ICO performance. Regarding the incremental informative value of white paper topics in predicting the issuer's post-ICO performance, the results notably contrast from the models on ICO success. In Panel A, which reports the results for the bootstrapped out-of-sample model performance, we find that for all models apart from Model 7, there is no significant improvement in average performance. These results indicate that the inclusion of the topic variables introduces model complexity to the extent that the average out-of-sample performance is no greater than the base model.

Furthermore, for individual topic categories, as shown in Panel B, we observe markedly fewer topic categories that significantly correlate with the post-ICO performance variables. For instance, apart from the topic category *Security*, which is negatively and significantly correlated with the token's first-day returns, first-month volatility and likelihood of delisting, few other categories are significant. Interestingly, other topic categories, such as *Blockchain*, *Network* and *Risk*, that were significant in explaining ICO success are no longer significant in explaining post-ICO performance.

Concerning the impact of the control variables, for $InitialReturns$, we find that none of the control variables are consequential. Regarding the $InitialVolatility$ in the first month, we find a negative influence of whitelisting/ KYC compliance. For $LongTermReturns$, projects in countries deemed as Tax havens, and those with restrictions on US investors and higher rating scores, observe more favorable outcome. Our results also show that projects with larger teams and lower topic diversity are less likely to get delisted.

Overall, our findings suggest that the influence of ICO white papers' thematic content subsides in the post-ICO period, as the project gains recognition and investors have access to more external information sources. After the ICO, the token listing marks an important event for investors. The market pricing mechanisms effectively start to aggregate all available information about the firm, and the token exchange listing and price discovery enable coordination among the investors (Momtaz, 2020a). For instance, a public release of a news on the firm's performance is reflected in the token price adjustments, along with all such information that may directly or indirectly influence firms' future performance. Such a coordination is not feasible during the ICO, when the token prices are static. Instead, investors must rely on active research, and often times the key details of the new project are limited to what is in the white paper. Furthermore, after a successful listing, and as the

project gains wider recognition and coverage, external and timely information are readily available to investors, making the white paper's content less relevant in their investment decisions, and thus, on the token's price action. Therefore, in contrast to the severe informational challenges during the ICO process, the information asymmetry decreases after the listing, making the content of the white papers less informative.

5.5. Additional analyses – ICO regulation and rating

Our results above are among the first to map the thematic content of ICO white papers and empirically show how the topics discussed in a white paper influence ICO investors. We find that the content of white papers is informative in explaining the ICO outcome, which echoes the arguments that actors intending to access external capital markets have the incentive to voluntarily disclose valuable information (e.g., Crawford and Sobel, 1982; Gigler, 1994; Stocken, 2000). More importantly, we find that specific topic categories, such as information on blockchain technology or descriptions of the mining procedures, which requires a degree of specialized knowledge and are difficult to mimic, as credible and favorable information. However, white papers' informativeness subsides with time as the project gains recognition; thus, limiting its value in explaining the issuing firm's future performance.

However, apart from the topics' inherent attributes, the degree to which voluntary disclosures, such as white papers, mitigate resource misallocation is likely to be sensitive to external factors that lend crediblity to the disclosure. We identify two credibility-enhancing mechanisms. First, we follow Shrestha et al. (2021) and investigate the influence of ICO-specific regulation on the information contained in the white paper. Second, complementing Bourveau et al. (2021), Barth et al. (2021) and Lee et al. (2021), we examine the impact of external scrutiny from a new type of information intermediary, ICO analysts.

5.5.1. The thematic content of ICO white papers in regulated vs. unregulated countries

We are first interested in whether white paper informativeness differs between projects located in countries with ICO-specific regulations and those without. Regulations surrounding ICOs remain preliminary and are still evolving with substantial variability across countries (Nestarcova, 2018). For instance, authorities in countries including Switzerland and Singapore have introduced extensive regulations and guidelines concerning ICOs, significantly curtailing investors' concerns; whereas in

most countries, a clear regulatory position yet remains to be established (Shrestha et al., 2021). We exploit this heterogeneity in regulation to examine the association between the thematic content of the white paper and the outcome variables, conditional on the regulatory status of the project's country location.

In evaluating regulations' relationship with the white papers' informativeness, two conflicting notions arise. First, regulations' impact on the information value of white papers stands central given its importance in reducing uncertainty by bringing credibility to economic transactions (Whittington, 1993). Regulation renders firms subject to litigation and, as a result, provide a basis for the project's legitimacy (Sutinen and Kuperan, 1999; Chelli, Durocher, and Richard, 2014). This increased credibility facilitates the ICOs' success, reduces risks, and supports the development of ICOs in countries with high institutional strength (Shrestha et al., 2021; Huang et al., 2018). Alternatively, prior research on the role of disclosure in capital markets shows that managers disclose information to signal their future prospects and reduce information asymmetry (Spence, 1978; Verrecchia, 2001). As such, in a context of high information asymmetry and limited regulation, informative white papers can allow investors to develop expectations about the project's future prospects, influencing their trading decisions. One could, therefore, also expect that, in countries with little or no regulation, white papers act as a voluntary bonding devise and are more consequential.

To test this relationship empirically, we distinguish the ICOs in our sample as *Regulated* and *Unregulated* based on the country-level ICO regulation data from the study of Shrestha et al. (2021).[21] The issuing firms launched in countries with ICO-related regulations or guidelines at the time of the launch are categorized as *Regulated*, and ICOs in countries that are yet to provide a clear regulatory direction concerning ICOs are identified as *Unregulated*. We drop the ICOs in countries with ICO bans, namely Algeria, China, Morocco and South Korea, given the limited number of observations.[22] As shown in Figure 2, the topic composition in white papers from ICOs launched in regulated countries substantially differs from ICOs launched in unregulated countries. The average white paper length is notably longer in the former group; and in particular, the topic categories *Blockchain*, *Security*, *People* and *Product* are more pronounced.

< Insert Figure 3 about here >

[21] We update Shrestha et al. (2021)'s data to include more recent changes ICO regulations in a number of countries. The updated regulation data is provided in Table B1 of the online appendix.

[22] While the bans on ICOs should prevent their launch in these countries, we find a number of ICOs launched after the introduction of the ban (43 cases). This exhibits the difficulties in regulating these fundraising efforts given their disintermediated structures. Qualifying such countries as "regulated" does not qualitatively change our results.

In Part A and B of Table 6, we report the regression results on the relationship between white paper topics and our set of dependent variables for regulated and unregulated countries, respectively. In Panel A of Part A, we report the bootstrapped statistics concerning model improvement from the base model and find results consistent with our main findings in Table 5. Likewise, in Panel B concerning the relationship with specific topic categories, the results remain largely consistent with our main findings, as topics *Blockchain*, *Mining*, *Product*, *Network* and *Innovation* are significant. However, we find contrasting evidence for unregulated countries. In Part B of Table 6, we find that in addition to reduced model performances, only two topics are associated with any of the ICO success variables. These findings indicate that investors are particularly skeptical of the white papers' content when the issuing project is based in an unregulated country.

< Insert Table 6 about here >

During the post-ICO period, however, we find that white papers from unregulated countries are more informative relative to those from regulated countries, particularly for the variable *InitialReturns*. We unite these findings under the umbrella of a reduced uncertainty after the moment of listing (Momtaz, 2020a). That is, before the token gets traded, investors face not only the problem of asymmetric information, but also the uncertainty pertaining to the protection of their investment. A lack of regulations will, therefore, limit the investor's trust in the content of the white paper. It is only once the token receives a stamp of approval by being issued and listed that investors consider the project as trustworthy and start incorporating the information contained in white papers. This evidence not only indicates that regulation surrounding ICOs promotes white papers' credibility but also points towards an under-reaction to the content of ICO white papers in countries with limited regulatory clarity.

5.5.2. The impact of ICO ratings

We next extend our analyses to examine the influence of external scrutiny from information intermediaries, such as ICO analysts, on the content of white papers and their information value. ICO analysts are experts who voluntarily provide ratings on ICO issuing firms' prospects on rating platforms (Lee et al., 2021). Unlike traditional credit rating agencies that receive direct compensation from issuers, which could potentially lead to ratings shopping and hence ratings inflation (Bolton, Freixas, and Shapiro, 2012), ICO analysts are not compensated for their advice (Lee et al., 2021).

Their main incentive is to enhance their own reputation in the industry (Bourveau et al., 2021). Demonstrating their ability to identify successful projects, the expert, in turn, receives higher rating from the platform, gaining more platform visibility and a greater likelihood of being hired as an advisor for subsequent ICOs.

Taken together, ICO analysts are expected to act as information intermediaries that help external capital providers with their investment decisions, mitigating the problems of due diligence and information processing (Boreiko and Vidusso, 2019). In line with this reasoning, Lee et al. (2021) and Roosenboom et al. (2020) find that expert ratings help in predicting ICO success and post-ICO performance, and Bourveau et al. (2021) show that white papers are a better predictor of ICO success when an ICO is rated. Yet, there are increasing evidence that the ICO analyst market is not void of conflicting interests and opportunistic behaviors, casting doubt in the extent of reliability and effectiveness of such ratings. In fact, Barth et al. (2021) reports that, even among ICOs with an average rating in the top quartile, more than 50% fail. They find that ICO analysts tend to reciprocate favorable ratings for their own projects and conclude that, although ratings predict ICO success, it only does so imperfectly (see also, Rhue, 2021). Given the question of ICO ratings' reliability, whether favorable ratings improve white papers' informativeness is an empirical question, which we test below.

We first examine how the topics in the white papers differ between ICOs with a high rating and low rating. Figure 4 illustrates the differences in the average number of topics sentences between the two groups. There are substantial differences between the two groups as all white paper topic categories receive significantly more attention in higher-rated ICOs, on average. The difference is greatest for categories $Network$, $Product$ and $Profitability$.

< Insert Figure 4 about here >

In Part A and B of Table 7, we report the regression results for sub-samples of high- and low-rated ICOs. We find notable differences in the influence of white papers, particularly in terms of ICO success. For ICOs with high ratings, several topics, including $Blockchain$, $Product$, $Innovation$ and $Risk$, influence multiple ICO success variables, and we observe significant model improvement in all three models. However, the evidence for white papers' informativeness substantially weakens for ICOs with lower ratings. In addition, for post-ICO performance, we observe similar diminished relationships among low-rated projects, suggesting an overall poorer informational value of white papers from low-rated projects.

< Insert Table 7 about here >

These results suggest that white papers are particularly informative when the ICO receives higher ratings from analysts. We interpret our results as evidence that, despite the potential opportunistic behavior (Barth et al., 2021), ICO analysts provide effective evaluations, which enhance the credibility of ICOs' white paper content. Overall, we complement the findings of Barth et al. (2021), Lee et al. (2021) and Bourveau et al. (2021), and highlight the positive certification role played by information intermediaries, such as ICO analysts, in a market with severe informational constraints.

5.6. Robustness tests

Our main findings on the relationship between white paper content and ICO success answers our research questions. However, we bear in mind that the exact quantification of this effect depends on the measurement of thematic content, as well as the model specifications used. Therefore, we now test the robustness of our findings.

5.6.1. Topic aggregation

First, in our main analyses, we grouped the 30 identified topics into 10 broad topic clusters for ease of interpretation. To show that our results hold if we consider the topics individually, in Table 8, we report the relationship between our different outcome variables and each topic variable ($Topic_{k,t,}$). We show that the findings in Table 8 are qualitatively comparable to those in Table 5. In Panel A, for models on ICO success, we find that the model performance significantly increases once we include $Topic_{k,t}$ to Equation 1. In contrast and in coherence with our main findings, the out-of-sample performance of the models on post-ICO performance is significantly poorer with the topic variables. Furthermore, in Panel B, the topics that are consistently significant belong to the same categories that we identified before, namely ICO, $Mining$, $Blockchain$ and $Network$. The disaggregated results also reveal some new insights. Separating the $Innovation$ category into $R\&D$ and $Data\&AI$ shows that while the discussion on application of AI and sophisticated data tools is consistently significant, $R\&D$ alone is not. Similarly, while the ICO category is not significant as an aggregate topic when we consider the four constituting topics separately, we find that sentences concerning $BuyIn$ influence investors' decisions favorably. However, we should note that in interpreting disaggregated results we should take caution, given the degree of overlaps between similar topics, as illustrated in Figure 1a.

For instance, while $Platform Development$ and $Platforms\&Apps$ appear to impact initial volatility in opposite directions, the results may be misleading as the two topics are conceptually similar. To avoid such misinterpretations and to provide more stable outputs, our main results report clustered topic categories that combine semantically similar topics.

< Insert Table 8 about here >

5.6.2. Number of topics

We then examine if our results are influenced by the number of topics we select as input in the sentLDA topic modeling process. To test the impact of the number of selected topics, we consider 15 topics instead of 30 and use the resulting estimations in our regressions. As shown in Panel A of Table 9, we find the thematic content holds incremental information value to explain the variables relating to ICO outcome but do not help explain post-ICO performance. This result supports our main analyses. Looking at the estimated coefficients for individual topics in Panel B, we find notable similarities between our main results with 30 topics and that with 15. For instance, for the topics $Blockchain$ and $ConsensusMechanism$, we find consistent positive relationships with ICO success. Similarly, the topic $Profitability$ and $Risk$ is again negatively associated with ICO success variables. The reduced model performance for post-ICO models also align with our main findings. Nevertheless, it should be noted that there are differences when sentences are classified into a different number of topics. Some topics may be disaggregated and others may be bunched into a single topic, leading to some differences in interpretation. Despite the potential alterations, as shown in Table 9, we find substantial consistency in how sentLDA maps different topics across sentences. The estimated model with 15 topics is considerably similar to our main results, and we can draw similar inferences despite taking half the number of topics as input.

< Insert Table 9 about here >

5.6.3. Informativeness over time

Finally, we examine how the relationship between white paper topic content and the outcome variables have evolved over time. The ICO market has observed substantial shifts in market activity through the years. For instance, the year 2017 was pivotal for ICOs, as the market, for the first time, saw exponential growth in both numbers and volume. During this time, ICOs also started receiving

significant attention from the mainstream media and regulators (for a detailed illustration of the ICO market's development, see Howell et al., 2020). However, the ICO market growth that peaked in early 2018 has since begun to subside (Masiak, Block, Masiak, Neuenkirch, and Pielen, 2019). Therefore, we investigate how the white papers informativeness has evolved from the fledgling bull period of 2017 and 2018 to the period of consolidation in the years 2019 and 2020. To that end, we divide our sample into three distinct parts: 2017 and earlier, 2018, and 2019 and after. We report the results in Table 10. For parsimony, we only report the significant topics and their coefficients.

< Insert Table 10 about here >

We find that, while for each distinct time period the inclusion of white paper topic categories improves model performance, there are some differences in which categories are significant over time. In the pre-2018 period, only *Network* significantly helps explain *TokenTraded*, while *Blockchain* and *Risk* are significant in explaining the amount raised. However, during the 2018 period, we find that substantially more topics are significant and positively associated with ICO success, including the technical topics, such as *Blockchain*, *Mining* and *Innovation*. We also find discussions of topics *Network*, *Risk* and *Product* are detrimental. In the post-2018 period, *Network*, *Risk* and *Blockchain* explain *TimeToListing*, whereas *Profitability* and *Risk* help explain amount raised. While the results for the year 2018 particularly align with our main results in Table 5, the topic categories that appear across the periods, such as *Blockchain*, *Network*, and *Risk* are still consistent.

In contrast, for post-ICO performance models, we find that substantially fewer topic categories are significant, generally leading to poorer model performance from adding the topic category variables. These results again reflect our findings in Table 5. Specifically, for pre-2018 period, the topic category *Innovation* significantly associated with *InitialReturns*, while *Profitability* with *InitialVolatility*. In 2018, we find that no topic significantly explains *InitialReturns*, while a higher emphasis on *Security* is linked with reduced *InitialVolatility*. Furthermore, the topic category *Mining* is significantly associated with *LongTermReturns* and *Delisted*. In the post-2018 period, we find no significant relationship between topics and post-ICO performance. Nonetheless, it should be noted that the sample of ICOs issued in the later years is substantially smaller.

6. Conclusion

Disclosures are central in capital markets. They not only support investors' trust and participation, but also help sustain a market's efficiency (Leuz and Verrecchia, 2000; Botosan, 1997; Healy and Palepu, 2001). While prior literature mostly focuses on the importance of disclosures that are mandated in regulated markets (Bushee and Leuz, 2005; Brüggemann, Kaul, Leuz, and Werner, 2018), we contribute by focusing on the informativeness of voluntary disclosures in the new, decentralized and largely unregulated market of ICOs. The ICO market is characterized by limited and heterogeneous regulation and high information asymmetry, which raises questions about the role and informativeness of unverifiable voluntary disclosures for raising capital. While some evidence in prior research highlights the role of white papers in mitigating information asymmetry among investors (e.g., Howell et al., 2020; Amsden and Schweizer, 2018; Fisch, 2019), there are studies showing that the presence of white papers in itself does not help identify ICOs of high quality (e.g., Adhami et al., 2018). Given this backdrop, we echo the longstanding debate between voluntary versus mandatory disclosures (Healy and Palepu, 2001) and provide timely evidence on the role played by ICO white papers in this emerging market. Relying on advanced machine learning methods (Bao and Datta, 2014), we depart from prior literature and focus on *what* is contained in the white paper instead of the *how* the information is disclosed.

We define a comprehensive sample of 5,210 white papers between August 2015 and June 2020 and provide three sets of descriptive evidence on the informativeness of the thematic content of ICO white papers. First, we employ the sentence-based LDA topic modeling method introduced by Bao and Datta (2014) to simultaneously specify and quantify the topics contained in ICO white papers. We identify 30 topics and find that the topics regarding the ICO's blockchain application and platform development are the most discussed topics in the white paper, while the topics regarding legal features and risk management appear to be the least discussed. Other topics such as data management, artificial intelligence tools, decentralization and energy consumption are also discussed in the white papers.

Second, we find that the white paper's thematic content is informative in explaining ICO performance and helps identify successful ICOs. Yet, we observe that not all categories share the same degree of relevance. We find that investors appear to deem some types of information as important and credible, while others seem superfluous. In particular, the topic categories that relate to the

technical features of ICOs (blockchain- and mining-related topics) are significantly linked with the ICO's success.

Third, and in contrast with the importance of white papers in explaining ICOs' performance, we find that white papers' informativeness substantially diminishes after the token is listed, which indicates that after the token's exchange listing, market pricing mechanisms perform a coordination function that dispersed investors fail to perform on their own during the ICO. Finally, we find that credibility-enhancing mechanisms (i.e. regulation and ICO analysts) reinforce the information value of ICOs' white papers. Overall, our analyses show that, despite their voluntary nature, investors view ICO white papers as informative to predict ICOs' performance and that such information is viewed as more credible when the ICO market is regulated and when the ICO is scrutinized by external experts.

Similar to prior research on ICOs, it is important to note that our results are descriptive and illustrate associations rather than causal relationships. Nonetheless, as token offerings continue to diversify with the emergence of new markets, such as IEOs and STOs, our results are central to investors, academics and regulators alike to better comprehend the importance of voluntary disclosures and credibility-enhancing mechanisms in markets with limited regulation. As discussed by Chod and Lyandres (2021), in light of the agency problems, for ICOs to remain a legitimate alternative for financing entrepreneurial ventures, regulations that protect investors are needed. Based on the automated, replicable and reliable classification of topics using a machine learning-based method, we highlight how regulators can draw a better understanding of what topics are discussed in the black box that constitute white papers, what type of information decreases information asymmetry and propose a standardization framework for white papers' content structure that mitigates the issues of asymmetric information. Our evidence also informs investors and regulators on the role played by ICO analysts in such decentralized markets and that, despite the lack of regulations, the ICO market has naturally found alternative ways to facilitate its efficiency and functioning as an alternative capital market through, for instance, ICO analysts.

References

Adhami, S., G. Giudici, and S. Martinazzi (2018). Why do businesses go crypto? An empirical analysis of initial coin offerings. *Journal of Economics and Business 100*, 64–75.

Agrawal, A., C. Catalini, and A. Goldfarb (2014). Some simple economics of crowdfunding. *Innovation Policy and the Economy 14*(1), 63–97.

Amsden, R. and D. Schweizer (2018). Are blockchain crowdsales the new 'Gold Rush'? Success determinants of Initial Coin Offerings. *Working Paper, Available at SSRN 3163849*.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica 71*(2), 579–625.

Bakos, Y. and H. Halaburda (2018). The role of cryptographic tokens and icos in fostering platform adoption. *Working Paper, Available at SSRN 3207777*.

Bao, Y. and A. Datta (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science 60*(6), 1371–1391.

Barraza, B. (2019). The Worth of Words: How Technical White Papers Influence ICO Blockchain Funding. *MIS Quarterly Executive 18*(4), 281–285.

Barth, A., V. Laturnus, S. Mansouri, and A. F. Wagner (2021). ICO analysts. *Swiss Finance Institute Research Paper No. 21-26, Available at SSRN 3720758*.

Bellstam, G., S. Bhagat, and J. A. Cookson (2020). A text-based analysis of corporate innovation. *Management Science 67*(7), 3985–4642.

Benedetti, H. and L. Kostovetsky (2021). Digital tulips? Returns to investors in initial coin offerings. *Journal of Corporate Finance 66*, 101786.

Benveniste, L. M. and P. A. Spindt (1989). How investment bankers determine the offer price and allocation of new issues. *Journal of Financial Economics 24*(2), 343–361.

Beyer, A., D. A. Cohen, T. Z. Lys, and B. R. Walther (2010). The financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics 50*(2-3), 296–343.

Blaseg, D. (2018). Dynamics of voluntary disclosure in the unregulated market for initial coin offerings. *Working Paper, Available at SSRN 3207641*.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Bolton, P., X. Freixas, and J. Shapiro (2012). The credit ratings game. *The Journal of Finance 67*(1), 85–111.

Boreiko, D. and G. Vidusso (2019). New blockchain intermediaries: Do ICO rating websites do their job well? *The Journal of Alternative Investments 21*(4), 67–79.

Botosan, C. A. (1997). Disclosure level and the cost of equity capital. *Accounting review 72*(3), 323–349.

Bourveau, T., E. T. De George, A. Ellahie, and D. Macciocchi (2021). The role of disclosure and information intermediaries in an unregulated capital market: evidence from initial coin offerings. *Journal of Accounting Research*.

Brown, S. and S. A. Hillegeist (2007). How disclosure quality affects the level of information asymmetry. *Review of Accounting Studies 12*(2-3), 443–477.

Brüggemann, U., A. Kaul, C. Leuz, and I. M. Werner (2018). The twilight zone: OTC regulatory regimes and market quality. *The Review of Financial Studies 31*(3), 898–942.

Bushee, B. J. and C. Leuz (2005). Economic consequences of SEC disclosure regulation: Evidence from the OTC bulletin board. *Journal of accounting and economics 39*(2), 233–264.

Bushway, S., B. D. Johnson, and L. A. Slocum (2007). Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of quantitative criminology 23*(2), 151–178.

Campbell, J. L., H. Chen, D. S. Dhaliwal, H.-m. Lu, and L. B. Steele (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies 19*(1), 396–455.

Campbell, K. and A. Minguez-Vera (2008). Gender diversity in the boardroom and firm financial performance. *Journal of business ethics 83*(3), 435–451.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pp. 288–296.

Chelli, M., S. Durocher, and J. Richard (2014). France's new economic regulations: Insights from institutional legitimacy theory. *Accounting, Auditing Accountability Journal 27*(2), 283–316.

Chen, Y. and C. Bellavitis (2020). Blockchain disruption and decentralized finance: The rise of decentralized business models. *Journal of Business Venturing Insights 13*, e00151.

Chod, J. and E. Lyandres (2021). A theory of icos: Diversification, agency, and information asymmetry. *Management Science 67*(10), 5969–5989.

Cohney, S., D. Hoffman, J. Sklaroff, and D. Wishnick (2019). Coin-operated capitalism. *Columbia Law Review 119*(3), 591–676.

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, 1431–1451.

Curme, C., T. Preis, H. E. Stanley, and H. S. Moat (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences 111*(32), 11600–11605.

Dittmar, R. F. and D. A. Wu (2019). Initial coin offerings hyped and dehyped: An empirical examination. *Working Paper, Available at SSRN 3259182*.

Dyer, T., M. Lang, and L. Stice-Lawrence (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics 64*(2-3), 221–245.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Eickhoff, M. and N. Neuss (2017). Topic modelling methodology: Its use in information systems and other managerial disciplines. In *25th European Conference on Information Systems (ECIS)*, pp. 1327–47.

El-Haj, M., P. Rayson, M. Walker, S. Young, and V. Simaki (2019). In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting 46*(3-4), 265–306.

Felix, T. H. and H. von Eije (2019). Underpricing in the cryptocurrency world: Evidence from initial coin offerings. *Managerial Finance 45*(4), 563–578.

Feng, C., N. Li, M. H. Wong, and M. Zhang (2019). Initial coin offerings, blockchain technology, and white paper disclosures. *Working Paper, Available at SSRN 3256289*.

Fisch, C. (2019). Initial coin offerings (ICOs) to finance new ventures. *Journal of Business Venturing 34*(1), 1–22.

Fisch, C., C. Masiak, S. Vismara, and J. Block (2019). Motives and profiles of ICO investors. *Journal of Business Research*, forthcoming.

Fisch, C. and P. P. Momtaz (2020). Institutional investors and post-ICO performance: An empirical analysis of investor returns in Initial Coin Offerings (ICOs). *Journal of Corporate Finance 64*(0929-1199), 101679.

Florysiak, D. and A. Schandlbauer (2021). The information content of ICO white papers. *Working Paper, Available at SSRN 3265007*.

Fu, C., A. Koh, and P. Griffin (2019). Automated theme search in ICO whitepapers. *The Journal of Financial Data Science 1*(4), 140–158.

Gigler, F. (1994). Self-enforcing voluntary disclosures. *Journal of Accounting Research 32*(2), 224–240.

Giorgi, S. and K. Weber (2015). Marks of distinction: Framing and audience appreciation in the context of investment advice. *Administrative Science Quarterly 60*(2), 333–367.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*(3), 267–297.

Grossman, S. J. and O. D. Hart (1980). Disclosure laws and takeover bids. *The Journal of Finance 35*(2), 323–334.

Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*(1), 29–36.

Harrell Jr, F. E., K. L. Lee, and D. B. Mark (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine 15*(4), 361–387.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science Business Media.

Healy, P. M. and K. G. Palepu (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics 31*(1-3), 405–440.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Hines, J. R. (2010). Treasure islands. *Journal of Economic Perspectives 24*(4), 103–126.

Hoberg, G. and C. Lewis (2017). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance 43*, 58–85.

Howell, S. T., M. Niessner, and D. Yermack (2020). Initial Coin Offerings: Financing growth with cryptocurrency token sales. *The Review of Financial Studies 33*(9), 3925–3974.

Huang, A. H., R. Lehavy, A. Y. Zang, and R. Zheng (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science 64*(6), 2833–2855.

Ivers, M. (2010). *Random House Guide to Good Writing*. Ballantine Books.

Janes, H., G. Longton, and M. S. Pepe (2009). Accommodating covariates in receiver operating characteristic analysis. *The Stata Journal 9*(1), 17–39.

Kaplan, S. and K. Vakili (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal 36*(10), 1435–1457.

Katsiampa, P. (2019). Volatility co-movement between Bitcoin and Ether. *Finance Research Letters 30*, 221–227.

Kaufmann, D., A. Kraay, and M. Mastruzzi (2010). The worldwide governance indicators: A summary of methodology. *Data and Analytical Issues, World Bank Policy Research Working Paper 5430*.

Kim, I., S. Miller, H. Wan, and B. Wang (2016). Drivers behind the monitoring effectiveness of global institutional investors: Evidence from earnings management. *Journal of Corporate Finance 40*, 24–46.

Kremer, I., Y. Mansour, and M. Perry (2014). Implementing the "wisdom of the crowd". *Journal of Political Economy 122*(5), 988–1012.

Lee, J., T. Li, and D. Shin (2021). The wisdom of crowds in FinTech: Evidence from initial coin offerings. *Available at SSRN 3195877*.

Leuz, C. and R. E. Verrecchia (2000). The economic consequences of increased disclosure. *Journal of Accounting Research*, 91–124.

Lewis, C. and S. Young (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research 49*(5), 587–615.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance 66*(1), 35–65.

Loughran, T. and B. McDonald (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research 54*(4), 1187–1230.

Lyandres, E., B. Palazzo, and D. Rabetti (2020). ICO success and post-ICO performance. *Working Paper, Available at SSRN 3287583*.

Masiak, C., J. H. Block, T. Masiak, M. Neuenkirch, and K. N. Pielen (2019). Initial coin offerings (ICOs): Market cycles and relationship with bitcoin and ether. *Small Business Economics*, 1–18.

Miglo, A. (2020). Choice Between IEO and ICO: Speed vs. Liquidity vs. Risk. *Working Paper, Available at SSRN 3561439*.

Miglo, A. (2021). STO vs. ICO: a theory of token issues under moral hazard and demand uncertainty. *Journal of Risk and Financial Management 14*(6), 232.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics 12*(2), 380–391.

Momtaz, P. P. (2019). Token sales and initial coin offerings: introduction. *The Journal of Alternative Investments 21*(4), 7–12.

Momtaz, P. P. (2020a). Entrepreneurial finance and moral hazard: Evidence from token offerings. *Journal of Business Venturing 36*(5), 106001.

Momtaz, P. P. (2020b). Initial coin offerings, asymmetric information, and loyal CEOs. *Small Business Economics*, forthcoming.

Momtaz, P. P. (2020c). The pricing and performance of cryptocurrency. *The European Journal of Finance 27*(4-5), 367–380.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages limitations. *Journal of Management 20*(4), 903–931.

Nestarcova, D. (2018). A critical appraisal of Initial Coin Offerings: Lifting the "Digital Token's Veil". *Brill Research Perspectives in International Banking and Securities Law 3*(2-3), 1–171.

Pencina, M. J. and R. B. D'Agostino (2004). Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Statistics in Medicine 23*(13), 2109–2123.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys 14*(1), 53–68.

Qiao, X., H. Zhu, and L. Hau (2020). Time-frequency co-movement of cryptocurrency return and volatility: Evidence from wavelet coherence analysis. *International Review of Financial Analysis 71*, 101541.

Reguera-Alvarado, N., P. de Fuentes, and J. Laffarga (2017). Does board gender diversity influence financial performance? Evidence from Spain. *Journal of Business Ethics 141*(2), 337–350.

Rhue, L. (2021). Trust is all you need: An empirical exploration of initial coin offerings (ICOs) and ICO reputation scores. *Journal of Insurance and Financial Management 4*(5), 44–79.

Roosenboom, P., T. van der Kolk, and A. de Jong (2020). What determines success in initial coin offerings? *Venture Capital 22*(2), 161–183.

Samieifar, S. and D. G. Baur (2020). Read me if you can! An analysis of ICO white papers. *Finance Research Letters*, 101427.

Sapkota, N. and K. Grobys (2021). Fear Sells: Determinants of Fund-Raising Success in the cross-section of Initial Coin Offerings. *Working Paper, Available at SSRN 3843138*.

Shannon, C. E. and W. Weaver (1963). The Mathematical Theory of Communication, Univ. of Ill. *Press, Urbana IL*.

Shrestha, P., Ö. Arslan-Ayaydin, J. Thewissen, and W. Torsin (2021). Institutions, Regulations, and Initial Coin Offerings: An International Perspective. *International Review of Economics and Finance 72*, 102–120.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics*, pp. 281–306. Elsevier.

Stocken, P. C. (2000). Credibility of voluntary disclosure. *The RAND Journal of Economics*, 359–374.

Sutinen, J. G. and K. Kuperan (1999). A socio-economic theory of regulatory compliance. *International Journal of Social Economics 26*(1/2/3), 174–193.

Tasca, P., P. Cerchiello, and A. M. Toma (2019). ICOs success drivers: A textual and statistical analysis. *The Journal of Alternative Investments 21*(4), 13–25.

Tiwari, M., A. Gepp, and K. Kumar (2019). The future of raising finance-a new opportunity to commit fraud: A review of initial coin offering (ICOs) scams. *Crime, Law and Social Change*, 1–25.

Tsamardinos, I., E. Greasidou, and G. Borboudakis (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning 107*(12), 1895–1922.

Verrecchia, R. E. (2001). Essays on disclosure. *Journal of Accounting and Economics 32*(1-3), 97–180.

Whittington, G. (1993). Corporate governance and the regulation of financial reporting. *Accounting and Business Research 23*(sup1), 311–319.

Zetzsche, D. A., R. P. Buckley, D. W. Arner, and L. Föhr (2017). The ICO Gold Rush: It's a scam, it's a bubble, it's a super challenge for regulators. *University of Luxembourg Law Working Paper* (11), 17–83.

Zhang, S., W. Aerts, L. Lu, and H. Pan (2019). Readability of token whitepaper and ICO first-day return. *Economics Letters 180*, 58–61.

Zhang, S., W. Aerts, D. Zhang, and Z. Chen (2021). Positive tone and initial coin offering. *Accounting & Finance*.

Zhang, Y., R. Jin, and Z.-H. Zhou (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics 1*(1-4), 43–52.

7. Tables

Table 1.: **Variable description**

Dependent Variables	
TokenTraded	Indicates whether the token is eventually traded on a currency exchange (Coinmarketcap).
Amount*	The amount raised during the coin-offering period in US dollars.
TimeToListing	The number of days from the end of the ICO to the listing of the issued token on an exchange platform.
InitialReturns	Market-adjusted price returns on the first day the token is traded.
InitialVolatility	Realized volatility in the token price during the first 30 days the token is traded.
LongTermReturns	Market-adjusted buy-and-hold price returns during the first 180 days since the token starts trading.
Delisted	Indicates if the token is delisted from Coinmarketcap.
Control Variables	
WP_Readability	Gunning Fog Index Readability score of the white paper text.
WP_Sentiment	Proportion of positive minus negative words in white paper text based on the dictionaries provided by Loughran and McDonald (2011).
WP_Pages	Number of pages in the ICO white paper PDF.
WP_TopicDiversity	Diversity in white paper's topic composition based on Shannon diversity index (Shannon and Weaver, 1963).
TaxHaven	Indicates whether the country is located in a tax haven (Hines, 2010).
Institutions	Aggregated institution score based on Worldwide Governance Indicators (Kaufmann et al., 2010).
USRestrict	Indicates if US-based investors are restricted from participating in the ICO.
Ratings	Average of the source-specific ratings in standardized values.
SocialMedia	Number of different social media channels used by the ICO project.
Video	Indicates whether the project provided a descriptive video.
Eth	Indicates whether the project blockchain is built on the Ethereum platform.
WhitelistKYC	Indicates whether the ICO implements Whitelisting and Know Your Customer (KYC) compliances.
Team	The number of members in the team behind the ICO.
TokenDist	Indicates whether the token distribution structure is specified.
MinInvest	Indicates whether a minimum investment amount is specified.
NumbCurr	The number of types of fiat and cryptocurrencies that the ICO accepts.
Fiat	Indicates whether the ICO accepts fiat currencies.
PreICO	Indicates whether a pre-ICO sale is conducted.
Hardcap	Indicates whether a soft cap is specified.
Softcap	Indicates whether a hard cap is specified.
Bonus	Indicates if a bonus scheme was offered to investors during the ICO.

Note: * indicates that natural logarithmic values are used in the regression models.

Table 2.: **Summary statistics of ICO sample**

Panel A: Summary Statistics of Dependent and Control Variables

	mean	sd	min	med	max
Dependent Variables					
(1) TokenTraded#	0.186	0.390	0	0	1
(2) Amount†	USD 8.85 mil	USD 12.31 mil	USD 1,080	USD 3.79 mil	USD 94.07
(3) TimeToListing◇	229.589	273.789	1	116	1,213
(4) InitialReturns‡	0.140	0.465	−0.385	0.028	5.422
(5) InitialVolatility§	0.844	0.495	0.164	0.704	3.284
(6) LongTermReturns¶	−0.403	2.771	−5.985	−0.458	26.794
(7) Delisted◇	0.350	0.477	0.000	0.000	1.000
Independent Variables					
(8) WP_Readability	15.589	2.123	7.772	15.572	22.912
(9) WP_Sentiment	0.002	0.006	−0.025	0.003	0.030
(10) WP_Pages	32.933	16.659	3	30	167
(11) WP_TopicDiversity	2.234	0.393	0.000	2.304	3.030
(12) TaxHaven	0.339	0.473	0	0	1
(13) Institutions	2.386	1.845	−4.326	3.081	4.448
(14) USRestrict	0.394	0.489	0	0	1
(15) Rating	0.132	0.772	−2.346	0.171	2.218
(16) SocialMedia	6.419	2.086	0	7	12
(17) Video	0.799	0.401	0	1	1
(18) Eth	0.855	0.352	0	1	1
(19) WhitelistKYC	0.659	0.474	0	1	1
(20) Team	12.414	7.512	1	11	69
(21) TokenDist	0.944	0.230	0	1	1
(22) MinInvest	0.434	0.496	0	0	1
(23) NumbCurr	2.251	1.826	1	2	30
(24) Fiat	0.242	0.428	0	0	1
(25) PreICO	0.687	0.464	0	1	1
(26) HardCap	0.856	0.351	0	1	1
(27) SoftCap	0.744	0.437	0	1	1
(28) Bonus	0.657	0.475	0	1	1

Panel B: Correlation Table

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	
(1) TokenTraded#														
(2) Amount†	0.31***													
(3) TimeToListing◇	-0.80***	-0.31***												
(4) InitialReturns‡	-	-0.06	0.05											
(5) InitialVolatility§	-	-0.15***	0.01	0.12**										
(6) LongTermReturns¶	-	0.00	-0.10**	0.05	0.07									
(7) Delisted◇	-	-0.09*	0.00	0.21***	0.08*	0.01								
(8) WP_Readability	0.00	0.04	-0.02	0.05	-0.06	-0.03	0.01							
(9) WP_Sentiment	-0.05**	-0.05*	0.06***	0.04	0.08*	0.02	0.00	-0.05***						
(10) WP_Pages	0.07***	0.07**	-0.12***	0.05	-0.05	-0.01	-0.10**	0.15***	-0.07***					
(11) WP_TopicDiversity	-0.02	-0.02	0.04*	0.00	0.01	-0.08	0.07*	0.16***	-0.04**	0.27***				
(12) TaxHaven	0.10***	0.09***	-0.11***	0.01	-0.17***	0.05	-0.07*	0.07***	-0.04**	0.13***	0.09***			
(13) Institutions	0.06***	0.11***	-0.07***	-0.05	-0.10**	-0.05	-0.07*	0.05**	-0.05***	0.08***	0.04**	0.31***		
(14) USRestrict	0.05**	-0.01	-0.07***	0.00	-0.11**	0.06	-0.07*	0.05***	-0.04**	0.14***	0.08***	0.14***	0.08***	
(15) Rating	0.25***	0.07**	-0.28***	0.09*	-0.06	0.08	-0.09**	0.04**	-0.02	0.27***	0.10***	0.10***	0.08***	
(16) SocialMedia	0.12***	-0.05*	-0.12***	0.07	-0.01	0.09*	-0.01	-0.01	-0.03*	0.18***	0.07***	0.04**	0.01	
(17) Video	0.12***	0.04	-0.10***	0.04	-0.01	-0.01	-0.05	0.04**	0.01	0.15***	0.05**	0.06***	0.03	
(18) Eth	0.03	-0.02	-0.04*	0.00	0.03	0.06	0.04	-0.02	-0.07***	0.06***	0.02	0.05**	0.03	
(19) WhitelistKYC	0.05***	0.06**	-0.12***	-0.03	-0.16***	0.00	-0.12***	0.09***	-0.02	0.22***	0.11***	0.14***	0.16***	
(20) Team	0.18***	0.12***	-0.15***	0.00	-0.04	-0.01	-0.08*	0.10***	0.02	0.36***	0.13***	0.13***	0.09***	
(21) TokenDist	0.06***	0.00	-0.10***	0.00	-0.03	0.07	-0.02	-0.02	-0.03*	0.03	0.07***	0.02	0.01	
(22) MinInvest	0.04**	-0.02	-0.07***	0.02	-0.11**	0.01	-0.08*	0.00	-0.05***	0.10***	0.03*	0.04*	0.02	
(23) NumbCurr	0.01	0.03	-0.04**	0.03	0.04	-0.07	-0.04	0.03	0.05**	0.09***	0.06***	-0.03*	-0.03*	
(24) Fiat	0.00	0.03	-0.08***	0.01	-0.03	-0.08	0.06	0.05***	0.04**	0.14***	0.07***	-0.01	0.04**	
(25) PreICO	-0.06***	-0.01	0.04**	0.02	-0.04	0.03	0.00	0.01	-0.01	0.11***	0.10***	0.02	-0.02	
(26) HardCap	0.07***	0.01	-0.10***	0.07	0.00	0.13***	0.02	0.03*	-0.04**	0.11***	0.11***	0.07***	0.05***	
(27) SoftCap	0.06***	-0.01	-0.10***	-0.05	-0.04	-0.01	-0.07	-0.01	-0.04**	0.09***	0.04*	0.05***	0.01	
(28) Bonus	-0.02	-0.03	0.06***	0.05	0.10**	0.08*	0.05	-0.04**	0.00	0.09***	0.10***	-0.03	0.00	

	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	
(15) Rating	0.24***													
(16) SocialMedia	0.27***	0.59***												
(17) Video	0.14***	0.35***	0.31***											
(18) Eth	0.08***	0.10***	0.14***	0.06***										
(19) WhitelistKYC	0.28***	0.32***	0.24***	0.18***	0.06***									
(20) Team	0.17***	0.38***	0.28***	0.21***	0.03*	0.22***								
(21) TokenDist	0.11***	0.20***	0.19***	0.07***	0.10***	0.12***	0.10***							
(22) MinInvest	0.23***	0.21***	0.24***	0.10***	0.12***	0.16***	0.10***	0.13***						
(23) NumbCurr	0.01	0.14***	0.10***	0.07***	-0.14***	0.06***	0.06***	0.05**	0.05***					
(24) Fiat	0.04*	0.13***	0.09***	0.08***	-0.01	0.13***	0.10***	0.03	0.07***	0.38***				
(25) PreICO	0.15***	0.18***	0.21***	0.10***	0.09***	0.15***	0.13***	0.15***	0.11***	0.05***	0.04*			
(26) HardCap	0.19***	0.29***	0.25***	0.14***	0.13***	0.23***	0.13***	0.23***	0.21***	0.07***	0.07***	0.14***		
(27) SoftCap	0.18***	0.27***	0.25***	0.14***	0.14***	0.20***	0.12***	0.18***	0.23***	0.08***	0.07***	0.12***	0.49***	
(28) Bonus	0.18***	0.17***	0.22***	0.11***	0.06***	0.11***	0.12***	0.19***	0.15***	0.12***	0.03	0.28***	0.17***	0.17***

Note: This table presents the summary statistics (mean, standard deviation, minimum, median and maximum), and the correlation matrix between the dependent variables and the control variables. The table shows Pearson correlation coefficients with significance levels 10 percent, 5 percent and 1 percent denoted with *, ** and ***, respectively.

Sample Overview
Total sample with white papers: 2,505
†Sample with amount raised: 1,203
◇ Sample of listed tokens with listing dates: 369
◇◇ Sample of listed tokens with listing dates and unlisted tokens: 2,403
‡Sample with token price data and listing dates: 369
§Sample with token price data for at least 30 days: 369
¶Sample with token price data for at least 6 months: 345
∓ Sample of listed tokens: 466

Table 3.: **White paper topic wordclouds**

(1) InitialSale

(2) BuyIn

(3) Liquidity

(4) Roadmap

(5) BlockchainApplication

(6) SmartContract

(7) ConsensusMechanism

(8) Governance

(9) EnergySustainability

(10) BlockchainEncryption

(11) Regulations

(12) Expertise

(13) HumanResource

(14) ServiceProfile

(15) FinancialServices

(16) Gaming

(17) Health

(18) Investment

(19) TokenBenefits

(20) TargetMarket

(21) MarketSize

(22) PlatformDevelopment

(23) Publishing

(24) Platforms&Apps

(25) R&D

(26) Data&AI

(27) RiskDisclosure

(28) RiskManagement

(29) LegalDisclaimers

(30) Terms&Conditions

Table 4.: **Topic description**

White Paper Topic	Description	mean	sd	min	med	max
ICO		**46.807**	**53.599**	**0**	**34**	**1,796**
(1) InitialSale	Various details relating to the ICO, such as duration and distribution	11.417	16.574	0	8	754
(2) BuyIn	Concerning various payment options for investors	11.692	24.452	0	3	636
(3) Liquidity	Information on the liquidity of the issued tokens and the secondary exchanges	12.441	29.898	0	4	1,207
(4) Roadmap	Project's future project and business goals	11.257	14.697	0	7	191
Blockchain		**27.419**	**34.460**	**0**	**17**	**483**
(5) BlockchainApplication	Application of blockchain for project implementation	16.800	22.940	0	9	321
(6) SmartContract	Application of smart contracts to deliver the said functionalities	10.619	19.968	0	4	394
Mining		**29.046**	**62.413**	**0**	**6**	**1,032**
(7) ConsensusMechanism	Underlying blockchain validation and consensus mechanism	13.030	45.065	0	1	1,028
(8) Governance	Governance system and protocols dictating the blockchain	8.534	24.521	0	0	453
(9) EnergySustainability	Energy consumption and sustainability issues	7.481	34.270	0	0	663
Security		**21.441**	**35.828**	**0**	**9**	**775**
(10) BlockchainEncryption	Encryption tools used in the blockchain	13.195	28.028	0	4	366
(11) Regulations	Regulatory oversights and due-diligence measures, such as KYC and whitelists	8.246	19.254	0	2	724
People		**19.978**	**27.700**	**0**	**8**	**259**
(12) Expertise	Specific expertise and qualification of the team members	9.174	17.157	0	2	223
(13) HumanResource	People involved and the nature of the job	10.804	15.341	0	4	133
Product		**38.600**	**62.549**	**0**	**12**	**1,168**
(14) ServiceProfile	General illustration of service and value proposition	10.222	34.333	0	0	1,168
(15) FinancialServices	Details relating to financial services and products	10.515	27.946	0	1	531
(16) Gaming	Specific application concerning gaming industry	11.285	35.444	0	0	557
(17) Health	Specific application concerning health industry	6.578	33.904	0	0	533
Profitability		**38.874**	**51.460**	**0**	**25**	**1,262**
(18) Investment	Projections of returns on investment	9.607	31.761	0	0	799
(19) TokenBenefits	Token utility and rewards systems to incentivize participation	11.632	25.358	0	6	1,258
(20) TargetMarket	Specific target market of the project	8.743	25.370	0	1	571
(21) MarketSize	Projected market size	8.892	13.533	0	4	221
Network		**37.085**	**52.738**	**0**	**21**	**1,881**
(22) PlatformDevelopment	On decentralization and community growth	15.543	26.052	0	6	318
(23) Publishing	Content generation and social media activity, mainly for the purpose of marketing	11.969	35.910	0	2	1,569
(24) Platforms&Apps	Concerning platform-based services and mobile applications	9.572	16.021	0	4	221
Innovation		**19.054**	**45.863**	**0**	**3**	**826**
(25) R&D	Details on project development mainly concerning research and innovation	5.844	26.728	0	0	826
(26) Data&AI	Data management and application of artificial intelligence tools	13.210	37.561	0	1	744
Risk		**19.467**	**43.356**	**0**	**8**	**1,449**
(27) RiskDisclosure	Disclosure of downside risks	2.448	14.750	0	0	825
(28) RiskManagement	Risk information and steps taken to mitigate them	7.002	20.006	0	2	895
(29) LegalDisclaimers	Legal statements and disclaimers	4.730	11.982	0	1	421
(30) Terms&Conditions	Details on various terms and conditions, including the rights of the investors	5.287	10.535	0	1	325
TotalSentences		297.771	216.425	2	260	6,224
UniqueTopics		18.644	4.300	1	19	29

Note: This table presents brief descriptions of the 30 identified white paper topics, along with their summary statistics (mean, standard deviation, minimum, median and maximum). The topics are further grouped into 10 semantic clusters for ease of interpretation. Our sample is composed of 2,505 white papers.

Table 5.: **White paper topics and ICO outcome**

	ICO Success			Post-ICO Performance			
	TokenTraded Logit Model 1	(log) *Amount* OLS Model 2	*TimeToListing* Cox P. Hazard Model 3	*InitialReturns* OLS (2nd Stage) Model 4	*InitialVolatility* OLS (2nd Stage) Model 5	*LongTermReturns* OLS (2nd Stage) Model 6	*Delisted* Logit (2nd Stage) Model 7
Panel A: Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.723	0.107	0.729	0.004	0.037	0.009	0.533
Models with Topic Variables (Eqn 2)	0.729	0.112	0.746	0.012	0.039	0.010	0.555
Difference	**0.012*****	**0.005*****	**0.017*****	-0.001	0.002	0.001	**0.022*****
Panel B: 10 Topic Categories (Intercept)	13.829	**13.144*****		**−0.604***	**1.195*****	**−4.355****	**−16.240*****
		(0.555)		(0.332)	(0.260)	(2.189)	(1.824)
Control Variables							
WP_Readability	−0.026	**0.087*****	−0.037	−0.001	0.009	−0.052	0.016
	(0.026)	(0.028)	(0.025)	(0.016)	(0.014)	(0.092)	(0.054)
WP_Sentiment	**−20.000****	−0.634	−12.567	1.763	3.046	12.378	0.909
	(9.668)	(8.929)	(10.149)	(4.603)	(4.133)	(27.011)	(16.966)
WP_Pages	0.002	0.005	−0.003	0.005	0.003	−0.022	−0.001
	(0.005)	(0.004)	(0.005)	(0.004)	(0.002)	(0.014)	(0.011)
WP_TopicDiversity	−0.200	−0.090	−0.238	−0.068	−0.052	−0.365	**0.489****
	(0.223)	(0.129)	(0.197)	(0.057)	(0.049)	(0.290)	(0.239)
TaxHaven	0.274	−0.097	0.178	−0.000	−0.139	**1.327***	0.198
	(0.209)	(0.173)	(0.173)	(0.058)	(0.106)	(0.702)	(0.367)
Institutions	0.044	**0.165*****	0.004	0.010	−0.001	−0.215	−0.054
	(0.051)	(0.037)	(0.042)	(0.014)	(0.020)	(0.168)	(0.088)
USRestrict	**0.363*****	0.117	0.151	−0.008	−0.002	**0.688****	−0.242
	(0.134)	(0.107)	(0.127)	(0.056)	(0.056)	(0.336)	(0.248)
Rating	**1.204*****	**0.491*****	**0.928*****	0.061	−0.022	**1.204***	−0.595
	(0.168)	(0.124)	(0.110)	(0.086)	(0.071)	(0.621)	(0.535)
SocialMedia	−0.035	**−0.111*****	**−0.057***	0.015	−0.007	−0.065	**0.118***
	(0.035)	(0.037)	(0.031)	(0.014)	(0.022)	(0.088)	(0.069)
Video	**0.534*****	0.116	**0.693*****	−0.034	0.067	0.656	−0.366
	(0.195)	(0.168)	(0.166)	(0.053)	(0.127)	(0.496)	(0.369)
Eth	−0.029	−0.081	0.216	−0.006	0.001	−0.051	0.283
	(0.161)	(0.210)	(0.164)	(0.055)	(0.045)	(0.596)	(0.537)
WhitelistKYC	0.162	**0.488*****	**−0.466*****	−0.004	**−0.128****	0.002	−0.285
	(0.133)	(0.130)	(0.130)	(0.055)	(0.053)	(0.326)	(0.384)
Team	**0.036*****	**0.034*****	**0.036*****	−0.001	0.005	−0.002	**−0.037****
	(0.012)	(0.006)	(0.010)	(0.003)	(0.003)	(0.021)	(0.017)
TokenDist	0.262	−0.252	0.629	0.051	0.246	1.472	−0.329
	(0.535)	(0.344)	(0.432)	(0.103)	(0.175)	(0.963)	(0.711)
ICO-Specific Control Variables (1st Stage)							
MinInvest	0.074	0.007	−0.075				
	(0.151)	(0.134)	(0.122)				
NumbCurr	−0.007	0.039	0.009				
	(0.045)	(0.037)	(0.033)				
Fiat	0.054	**0.407*****	−0.178				
	(0.116)	(0.135)	(0.128)				
PreICO	**−0.284****	−0.004	**−0.469*****				
	(0.129)	(0.147)	(0.098)				
HardCap	0.224	0.034	0.089				
	(0.225)	(0.209)	(0.219)				
SoftCap	0.022	0.052	−0.200				
	(0.204)	(0.189)	(0.165)				
Bonus	−0.204	−0.096	−0.043				
	(0.129)	(0.129)	(0.132)				
White Paper Topic Variables							
(1) ICO	−0.002	−0.000	0.000	−0.001	0.000	0.005	0.002
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)
(2) Blockchain	**0.005****	**0.005*****	**0.006*****	−0.000	−0.000	0.001	−0.005
	(0.002)	(0.001)	(0.001)	(0.000)	(0.000)	(0.003)	(0.004)
(3) Mining	**0.003****	0.001	**0.002***	−0.000	−0.000	**0.007****	**−0.007*****
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.003)	(0.002)
(4) Security	0.000	0.001	0.002	**−0.002****	**−0.002*****	0.008	**−0.009*****
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.007)	(0.003)
(5) People	0.002	0.001	0.001	0.000	**−0.003****	−0.003	0.003
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.005)	(0.004)
(6) Product	**−0.002***	−0.000	**−0.002****	−0.001	0.000	0.010	0.000
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.007)	(0.002)
(7) Profitability	0.001	0.001	0.000	−0.001	−0.000	0.004	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)
(8) Network	**−0.003****	**−0.003*****	**−0.002***	0.000	0.000	−0.001	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)
(9) Innovation	0.001	0.000	**0.002****	0.001	−0.000	−0.000	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.002)	(0.002)
(10) Risk	**−0.005***	−0.002	**−0.005****	0.000	−0.001	0.008	0.004
	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.009)	(0.004)
Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
Num. obs.	2505	1203	2403	369	369	345	466
McFadden/Adj./Nagelkerke R^2	0.254	0.198	0.124	−0.039	0.172	0.092	0.166
Panel C: 2 Topic Clusters							
Technical	**0.002*****	**0.002*****	**0.003*****	−0.000	−0.000	0.003	**−0.005*****
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.002)	(0.001)
NonTechnical	**−0.001****	−0.000	−0.001	−0.000	−0.000	0.004	0.001
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.003)	(0.001)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
Num. obs.	2505	1203	2403	369	369	345	466
McFadden/Adj./Nagelkerke R^2	0.248	0.191	0.117	−0.042	0.156	0.080	0.156

Note: The table presents the results of our main analysis. It includes results for models concerning ICO success and post-ICO performance, where the first three columns relate to models for *TokenTraded*, *Amount*, and *TimeToListing*, and the following four columns relate to models concerning *InitialReturns*, *InitialVolatility*, *LongTermReturns*, and *Delisted*. In Panel A, we report the results from the comparative out-of-sample tests of the prediction models with topic category variables (Eq. 2) and without (Eq. 1). The performance metrics AUC, R^2, and concordance index (for the logistic, linear, and hazard models, respectively) are produced using simulated random data bootstrapped with 1,000 replications. The statistical significance of the differences in test statistics is determined with non-parametric Wald tests. In Panel B, the estimated coefficients concerning the relationships between the various white paper topic categories and ICO success and post-ICO performance measures are provided. The color-coded boxes indicate the estimated significance and direction of each topic variable's coefficients, where a green box indicates a positive relationship, a red box indicates a negative relationship, and a grey box indicates no significant relationship. As goodness-of-fit measures, McFadden, Adjusted, and Nagelkerke R^2 are provided for the Logit, OLS, and Cox P. Hazard models, respectively. *, ** and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on two-sided t-tests. All variables are defined in Tables 1 and 4. Furthermore, Panel C provides regression results from models with aggregated *Technical* and *NonTechnical* topic explanatory variables.

Table 6.: **White paper topics and ICO outcome by country regulation**

Part A: ICOs in Regulated Countries

	ICO Success			Post-ICO Performance			
	TokenTraded Logit Model 1	(log) *Amount* OLS Model 2	*TimeToListing* Cox P. Hazard Model 3	*InitialReturns* OLS (2nd) Model 5	*InitialVolatility* OLS (2nd) Model 6	*LongTermReturns* OLS (2nd) Model 7	*Delisted* Logit (2nd) Model 8
Panel A: Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.740	0.082	0.759	0.013	0.026	0.023	0.503
Models with Topic Variables (Eqn 2)	0.744	0.091	0.778	0.010	0.033	0.014	0.536
Difference	**0.004*****	**0.008*****	**0.019*****	**−0.003*****	**0.007*****	**−0.009*****	**0.033*****
Panel B: 10 Topic Categories							
(1) ICO	−0.000	0.001	0.002	−0.001	0.000	0.004	−0.000
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
(2) Blockchain	0.003	**0.005*****	**0.006*****	−0.000	−0.000	−0.002	−0.004
	(0.002)	(0.002)	(0.001)	(0.000)	(0.000)	(0.003)	(0.003)
(3) Mining	0.002	**0.002***	0.002	0.000	0.000	0.005	**−0.009*****
	(0.002)	(0.001)	(0.001)	(0.000)	(0.001)	(0.004)	(0.003)
(4) Security	0.000	0.002	0.001	−0.001	**−0.002****	0.005	**−0.012*****
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.007)	(0.005)
(5) People	0.001	0.000	−0.001	0.000	−0.003	0.002	0.008
	(0.002)	(0.003)	(0.003)	(0.001)	(0.002)	(0.003)	(0.006)
(6) Product	**−0.003***	0.001	**−0.002****	−0.000	0.001	0.010	−0.000
	(0.002)	(0.001)	(0.001)	(0.000)	(0.001)	(0.007)	(0.003)
(7) Profitability	−0.000	**0.002****	−0.002	0.000	−0.000	0.003	0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
(8) Network	**−0.003****	**−0.004*****	−0.001	0.000	0.000	−0.001	−0.001
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.002)	(0.004)
(9) Innovation	0.002	−0.001	**0.003*****	0.001	−0.001	−0.001	−0.003
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.004)
(10) Risk	−0.005	−0.002	−0.004	0.002	−0.000	0.012	0.004
	(0.003)	(0.003)	(0.003)	(0.001)	(0.001)	(0.012)	(0.006)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
Num. obs.	1719	791	1646	262	262	242	331
McFadden/Adj./Nagelkerke R^2	0.263	0.179	0.121	0.157	0.194	0.215	0.198

Part B: ICOs in Unregulated Countries

	ICO Success			Post-ICO Performance			
	TokenTraded Logit Model 8	(log) *Amount* OLS Model 9	*TimeToListing* Cox P. Hazard Model 10	*InitialReturns* OLS (2nd) Model 11	*InitialVolatility* OLS (2nd) Model 12	*LongTermReturns* OLS (2nd) Model 13	*Delisted* Logit (2nd) Model 14
Panel A: Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.631	0.076	0.739	0.030	0.049	0.057	0.488
Models with Topic Variables (Eqn 2)	0.614	0.071	0.752	0.055	0.048	0.047	0.483
Difference	**−0.018*****	**−0.005*****	**0.013*****	**0.025*****	-0.001	-0.010	-0.005
Panel B: 10 Topic Categories							
(1) ICO	−0.005	**−0.007*****	−0.008	−0.003	0.000	0.011	**0.214***
	(0.006)	(0.003)	(0.005)	(0.005)	(0.002)	(0.014)	(0.114)
(2) Blockchain	0.003	0.005	0.006	**−0.017****	−0.002	0.017	−0.162
	(0.007)	(0.005)	(0.004)	(0.008)	(0.004)	(0.022)	(0.102)
(3) Mining	0.002	−0.001	0.004	−0.002	−0.001	0.008	0.037
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.010)	(0.023)
(4) Security	−0.001	−0.002	0.002	−0.003	−0.002	−0.021	0.075
	(0.008)	(0.007)	(0.008)	(0.007)	(0.003)	(0.022)	(0.053)
(5) People	0.004	0.005	0.005	−0.011	**−0.004****	−0.018	**−0.241****
	(0.006)	(0.005)	(0.004)	(0.011)	(0.002)	(0.033)	(0.117)
(6) Product	0.002	0.001	0.002	0.001	**−0.002***	−0.001	0.029
	(0.004)	(0.002)	(0.002)	(0.003)	(0.001)	(0.010)	(0.022)
(7) Profitability	0.005	0.001	**0.005*****	−0.006	−0.001	0.012	0.009
	(0.004)	(0.003)	(0.002)	(0.004)	(0.001)	(0.013)	(0.016)
(8) Network	−0.001	−0.001	−0.003	0.005	**−0.005*****	−0.006	−0.069
	(0.002)	(0.001)	(0.002)	(0.006)	(0.001)	(0.016)	(0.058)
(9) Innovation	0.003	0.002	0.001	−0.001	0.000	−0.000	**−0.081***
	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.003)	(0.042)
(10) Risk	−0.007	−0.007	−0.010	−0.015	−0.006	−0.023	**0.150****
	(0.007)	(0.005)	(0.008)	(0.012)	(0.005)	(0.023)	(0.067)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
Num. obs.	651	341	629	68	68	65	89
McFadden/Adj./Nagelkerke R^2	0.329	0.209	0.098	−0.044	0.525	−0.193	0.701

Note: Parts A and B of this table present the results obtained from sub-samples of ICOs issued by projects based in *Unregulated* and *Regulated* countries, respectively. Each part provides results for models concerning ICO success (*TokenTraded*, *Amount* and *TimeToListing*) and post-ICO performance (*InitialReturns*, *InitialVolatility*, *LongTermReturns* and *Delisted*). In Panel A, we report the results from the comparative out-of-sample tests of the prediction models with topic category variables (Eq. 2) and without (Eq. 1). The performance metrics AUC, R^2 and concordance index (for the logistic, linear and hazard models, respectively) are produced using simulated random data bootstrapped with 1,000 replications. The statistical significance of the differences in test statistics is determined with non-parametric Wald tests. In Panel B, the estimated coefficients concerning the relationships between the various white paper topic categories and ICO success and post-ICO performance measures are provided. The color-coded boxes indicate the estimated significance and direction of each topic variable's coefficients, where a green box indicates a positive relationship, a red box indicates a negative relationship, and a grey box indicates no significant relationship. As goodness-of-fit measures, McFadden, Adjusted and Nagelkerke R^2 are provided for the Logit, OLS and Cox P. Hazard models, respectively. *, ** and *** denote statistical significance at the 10 percent, 5 percent and 1 percent levels, respectively, based on two-sided t-tests. All variables are defined in Tables 1 and 4.

Table 7.: **White paper topics and ICO outcome by high and low rating**

Part A: ICOs with High Rating

	ICO Success			**Post-ICO Performance**			
	TokenTraded Logit Model 1	(log) *Amount* OLS Model 2	*TimeToListing* Cox P. Hazard Model 3	*InitialReturns* OLS (2nd) Model 5	*InitialVolatility* OLS (2nd) Model 6	*LongTermReturns* OLS (2nd) Model 7	*Delisted* Logit (2nd) Model 8
Panel A: Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.691	0.101	0.650	0.005	0.023	0.014	0.528
Models with Topic Variables (Eqn 2)	0.693	0.103	0.683	0.005	0.026	0.019	0.549
Difference	**0.002*****	**0.002*****	**0.033*****	0.000	**0.003*****	**0.005****	**0.021*****
Panel B: 10 Topic Categories							
(1) ICO	−0.002	−0.001	−0.000	−0.001	0.000	**0.008****	0.001
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)
(2) Blockchain	0.004	**0.004****	**0.006*****	−0.000	−0.000	−0.000	−0.002
	(0.003)	(0.002)	(0.001)	(0.000)	(0.001)	(0.005)	(0.005)
(3) Mining	0.001	0.001	0.002	**−0.001****	−0.000	**0.010****	**−0.011*****
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.004)	(0.004)
(4) Security	−0.001	0.002	0.001	**−0.002***	**−0.001****	0.014	**−0.011****
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.010)	(0.004)
(5) People	0.003	0.001	0.001	0.001	**−0.002***	−0.005	0.006
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.006)	(0.005)
(6) Product	**−0.003****	−0.000	**−0.002****	−0.001	0.001	**0.016****	−0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.008)	(0.003)
(7) Profitability	−0.001	0.001	0.001	**−0.002***	−0.000	0.003	**−0.005***
	(0.002)	(0.001)	(0.002)	(0.001)	(0.000)	(0.003)	(0.003)
(8) Network	−0.002	**−0.002****	−0.002	0.001	0.000	−0.000	−0.003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.004)
(9) Innovation	**0.003****	0.001	**0.003****	0.001	**−0.001****	0.001	−0.004
	(0.001)	(0.002)	(0.001)	(0.001)	(0.000)	(0.002)	(0.003)
(10) Risk	**−0.007*****	−0.003	**−0.006*****	0.000	−0.001	0.015	0.004
	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.013)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
Num. obs.	1252	757	1186	276	276	258	338
McFadden/Adj./Nagelkerke R^2	0.221	0.189	0.099	−0.050	0.138	0.123	0.219

Part B: ICOs with Low Rating

	ICO Success			**Post-ICO Performance**			
	TokenTraded Logit Model 8	(log) *Amount* OLS Model 9	*TimeToListing* Cox P. Hazard Model 10	*InitialReturns* OLS (2nd) Model 11	*InitialVolatility* OLS (2nd) Model 12	*LongTermReturns* OLS (2nd) Model 13	*Delisted* Logit (2nd) Model 14
Panel A: Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.668	0.050	0.725	0.036	0.061	0.073	0.525
Models with Topic Variables (Eqn 2)	0.661	0.047	0.702	0.056	0.051	0.061	0.513
Difference	**−0.007*****	**−0.003***	**−0.023*****	**0.020*****	**−0.010***	**−0.012***	-0.012
Panel B: 10 Topic Categories							
(1) ICO	−0.003	0.001	−0.001	−0.001	−0.001	0.006	−0.007
	(0.003)	(0.003)	(0.003)	(0.001)	(0.002)	(0.006)	(0.012)
(2) Blockchain	0.005	**0.007***	**0.007****	0.000	−0.001	0.003	−0.007
	(0.004)	(0.004)	(0.003)	(0.000)	(0.001)	(0.002)	(0.008)
(3) Mining	**0.005****	0.002	0.003	−0.000	−0.001	−0.000	0.002
	(0.002)	(0.002)	(0.003)	(0.000)	(0.001)	(0.003)	(0.011)
(4) Security	0.003	−0.004	**0.005***	−0.001	**−0.005****	0.004	−0.009
	(0.004)	(0.005)	(0.003)	(0.001)	(0.002)	(0.005)	(0.009)
(5) People	−0.000	0.001	0.002	−0.003	−0.003	−0.005	0.035
	(0.004)	(0.006)	(0.004)	(0.002)	(0.004)	(0.013)	(0.028)
(6) Product	0.001	−0.001	−0.001	0.000	−0.002	−0.002	**0.013****
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.004)	(0.007)
(7) Profitability	0.002	0.001	−0.001	0.001	**−0.003***	0.008	**0.031*****
	(0.002)	(0.003)	(0.003)	(0.001)	(0.002)	(0.006)	(0.011)
(8) Network	−0.004	**−0.005***	−0.002	−0.001	−0.002	−0.003	**0.020****
	(0.004)	(0.003)	(0.003)	(0.001)	(0.001)	(0.006)	(0.008)
(9) Innovation	−0.001	−0.001	−0.002	−0.001	0.001	0.002	0.015
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.006)	(0.009)
(10) Risk	−0.003	−0.001	−0.004	−0.003	0.004	**−0.017***	−0.006
	(0.003)	(0.005)	(0.005)	(0.002)	(0.004)	(0.010)	(0.016)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
Num. obs.	1253	446	1217	93	93	87	128
McFadden/Adj./Nagelkerke R^2	0.258	0.163	0.057	0.195	0.336	0.418	0.476

Note: Parts A and B of this table present the results obtained from sub-samples of ICOs with high (>median) and low (≤median) aggregate ratings, respectively. Each part provides results for models concerning ICO success (*TokenTraded*, *Amount* and *TimeToListing*) and post-ICO performance (*InitialReturns*, *InitialVolatility*, *LongTermReturns* and *Delisted*). In Panel A, we report the results from the comparative out-of-sample tests of the prediction models with topic category variables (Eq. 2) and without (Eq. 1). The performance metrics AUC, R^2 and concordance index (for the logistic, linear and hazard models, respectively) are produced using simulated random data bootstrapped with 1,000 replications. The statistical significance of the differences in test statistics is determined with non-parametric Wald tests. In Panel B, the estimated coefficients concerning the relationships between the various white paper topic categories and ICO success and post-ICO performance measures are provided. The color-coded boxes indicate the estimated significance and direction of each topic variable's coefficients, where a green box indicates a positive relationship, a red box indicates a negative relationship, and a grey box indicates no significant relationship. As goodness-of-fit measures, McFadden, Adjusted and Nagelkerke R^2 are provided for the Logit, OLS and Cox P. Hazard models, respectively. *, ** and *** denote statistical significance at the 10 percent, 5 percent and 1 percent levels, respectively, based on two-sided t-tests. All variables are defined in Tables 1 and 4.

Table 8.: **Sensitivity tests – individual topics**

		ICO Success			Post-ICO Performance			
		TokenTraded Logit Model 1	(log) *Amount* OLS Model 2	*TimeToListing* Cox P. Hazard Model 3	*InitialReturns* OLS (2nd Stage) Model 4	*InitialVolatility* OLS (2nd Stage) Model 5	*LongTermReturns* OLS (2nd Stage) Model 6	*Delisted* Logit (2nd Stage) Model 7
Panel A: Model Comparison		*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
	Base Model (Eqn 1)	0.761	0.116	0.688	0.005	0.041	0.012	0.537
	Models with Topic Variables (Eqn 2)	0.764	0.120	0.731	0.004	0.031	0.007	0.561
	Difference	**0.003*****	**0.004****	**0.043*****	**−0.001***	**−0.010*****	**−0.005*****	**0.024*****
Panel B: 10 Topic Categories								
ICO	(1) InitialSale	−0.004	0.004	0.000	−0.003	−0.001	−0.008	0.002
		(0.005)	(0.005)	(0.005)	(0.002)	(0.002)	(0.014)	(0.002)
	(2) BuyIn	**0.006****	0.001	**0.005****	0.000	−0.001	0.002	**0.002****
		(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.005)	(0.001)
	(3) Liquidity	**−0.007****	0.001	**−0.007****	0.001	0.000	0.000	0.002
		(0.003)	(0.002)	(0.003)	(0.001)	(0.002)	(0.006)	(0.001)
	(4) Roadmap	−0.001	−0.005	0.005	−0.002	0.001	0.019	**−0.004*****
		(0.006)	(0.003)	(0.005)	(0.002)	(0.002)	(0.013)	(0.001)
Blockchain	(5) BlockchainApplication	**0.004***	0.003	**0.005*****	0.001	−0.001	0.007	0.000
		(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.006)	(0.001)
	(6) SmartContract	0.004	**0.006*****	**0.006*****	−0.001	0.000	−0.003	−0.001
		(0.003)	(0.002)	(0.001)	(0.000)	(0.000)	(0.003)	(0.001)
Mining	(7) ConsensusMechanism	**0.005***	**0.003***	**0.002***	**−0.001*****	−0.000	0.002	**−0.001*****
		(0.002)	(0.002)	(0.001)	(0.000)	(0.001)	(0.002)	(0.000)
	(8) Governance	**0.006*****	0.000	**0.006*****	0.001	−0.001	0.008	−0.001
		(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.008)	(0.001)
Security	(9) EnergySustainability	−0.000	0.001	−0.000	−0.001	−0.000	0.003	−0.000
		(0.001)	(0.001)	(0.002)	(0.001)	(0.000)	(0.004)	(0.000)
	(10) BlockchainEncryption	−0.001	−0.000	0.001	**−0.001****	**−0.001****	0.002	**−0.001****
		(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.005)	(0.001)
	(11) Regulations	0.002	0.007	0.003	−0.002	−0.004	−0.011	−0.001
		(0.005)	(0.005)	(0.005)	(0.002)	(0.003)	(0.017)	(0.002)
People	(12) Expertise	0.003	0.003	0.004	−0.001	**−0.005***	−0.016	0.001
		(0.003)	(0.003)	(0.005)	(0.001)	(0.002)	(0.011)	(0.001)
	(13) HumanResource	0.004	−0.002	0.000	0.000	−0.001	0.009	0.000
		(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.009)	(0.001)
Product	(14) ServiceProfile	−0.000	0.001	−0.000	−0.000	0.001	0.014	0.000
		(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.011)	(0.000)
	(15) FinancialServices	**−0.005*****	0.000	−0.003	−0.001	0.001	0.000	−0.001
		(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.004)	(0.001)
	(16) Gaming	−0.002	0.002	−0.001	−0.001	0.000	−0.001	0.000
		(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.001)
	(17) Health	−0.001	0.001	−0.003	0.001	−0.000	0.013	0.000
		(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.016)	(0.001)
Profitability	(18) Investment	0.002	**0.003****	0.002	−0.001	−0.001	**0.008****	−0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)
	(19) TokenBenefits	0.002	0.001	−0.000	−0.001	−0.001	−0.006	0.001
		(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.006)	(0.001)
	(20) TargetMarket	−0.003	−0.002	−0.004	−0.001	0.000	0.001	0.000
		(0.003)	(0.003)	(0.004)	(0.001)	(0.002)	(0.004)	(0.001)
	(21) MarketSize	−0.000	0.002	−0.002	−0.002	−0.000	−0.001	−0.001
		(0.004)	(0.004)	(0.005)	(0.002)	(0.002)	(0.008)	(0.001)
Network	(22) PlatformDevelopment	**−0.008****	**−0.007*****	**−0.006****	0.001	0.003	0.002	0.001
		(0.003)	(0.002)	(0.003)	(0.001)	(0.002)	(0.007)	(0.001)
	(23) Publishing	−0.001	−0.001	−0.001	−0.000	−0.001	−0.002	0.001
		(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)
	(24) Platforms&Apps	−0.001	**0.006***	0.002	−0.001	−0.001	−0.000	**−0.006*****
		(0.004)	(0.003)	(0.003)	(0.001)	(0.001)	(0.009)	(0.001)
Innovation	(25) R&D	−0.002	−0.001	−0.001	0.002	0.001	−0.003	−0.000
		(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.011)	(0.001)
	(26) Data&AI	**0.004*****	**0.003***	**0.004*****	0.000	**−0.001****	−0.004	−0.000
		(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.000)
Risk	(27) RiskDisclosure	−0.015	0.003	−0.006	0.002	0.008	0.081	0.000
		(0.013)	(0.013)	(0.010)	(0.008)	(0.008)	(0.069)	(0.007)
	(28) RiskManagement	−0.001	−0.005	−0.006	0.005	−0.001	−0.019	0.002
		(0.006)	(0.006)	(0.008)	(0.004)	(0.003)	(0.025)	(0.002)
	(29) LegalDisclaimers	−0.012	**−0.009*****	−0.010	−0.002	0.002	0.025	−0.002
		(0.009)	(0.003)	(0.008)	(0.004)	(0.005)	(0.053)	(0.004)
	(30) Terms&Conditions	0.005	0.004	−0.004	−0.003	−0.005	−0.018	0.001
		(0.008)	(0.005)	(0.007)	(0.002)	(0.004)	(0.021)	(0.004)
	Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
	Num. obs.	2505	1203	2403	369	369	345	466
	McFadden/Adj./Nagelkerke R^2	0.269	0.200	0.138	−0.063	0.150	0.021	0.190

Note: The table provides the results for the sensitivity analysis with unclustered groups of topics derived from the main sentLDA topic modeling results with 30 topics. The table includes results for models concerning ICO success and post-ICO performance, where the first three columns relate to models for *TokenTraded*, *Amount* and *TimeToListing*, and the following four columns relate to models concerning *InitialReturns*, *InitialVolatility*, *LongTermReturns* and *Delisted*. In Panel A, we report the results from the comparative out-of-sample tests of the prediction models with topic variables (Eq. 2) and without (Eq. 1). The performance metrics AUC, R^2 and concordance index (for the logistic, linear and hazard models, respectively) are produced using simulated random data bootstrapped with 1,000 replications. The statistical significance of the differences in test statistics is determined with non-parametric Wald tests. In Panel B, the estimated coefficients concerning the relationships between the various white paper topics and ICO success and post-ICO performance measures are provided. The color-coded boxes indicate the estimated significance and direction of each topic variable's coefficients, where a green box indicates a positive relationship, a red box indicates a negative relationship, and a grey box indicates no significant relationship. As goodness-of-fit measures, McFadden, Adjusted and Nagelkerke R^2 are provided for the Logit, OLS and Cox P. Hazard models, respectively. *, ** and *** denote statistical significance at the 10 percent, 5 percent and 1 percent levels, respectively, based on two-sided t-tests. All variables are defined in Tables 1 and 4.

Table 9.: **Sensitivity tests – sentLDA for $K = 15$**

	ICO Success			**Post-ICO Performance**			
	TokenTraded Logit Model 1	(log) *Amount* OLS Model 2	*TimeToListing* Cox P. Hazard Model 3	*InitialReturns* OLS (2nd) Model 4	*InitialVolatility* OLS (2nd) Model 5	*LongTermReturns* OLS (2nd) Model 6	*Delisted* Logit Model 7
Panel A: Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.701	0.091	0.665	0.006	0.049	0.029	0.545
Models with Topic Variables (Eqn 2)	0.707	0.122	0.694	0.007	0.046	0.017	0.547
Difference	**0.006*****	**0.03*****	**0.029*****	0.001	**−0.02*****	**−0.013*****	0.03
Panel B: 10 Topic Categories							
(1) ICO	**−0.006***	−0.002	−0.001	−0.001	0.000	−0.005	−0.003
	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.008)	(0.007)
(2) Liquidity	**0.004****	**0.003****	0.003	−0.001	−0.000	0.003	0.004
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.003)
(3) Blockchain	**0.004***	**0.003***	**0.004****	0.000	0.001	−0.002	0.003
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)
(4) ConsensusMechanism	**0.005*****	**0.005*****	**0.006*****	−0.000	−0.000	0.006	−0.001
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.003)
(5) EnergySustainability	−0.002	−0.000	−0.003	−0.000	0.000	0.001	0.001
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.003)
(6) BlockchainEncryption	0.002	0.001	0.001	−0.000	−0.000	−0.000	**−0.008*****
	(0.002)	(0.001)	(0.001)	(0.000)	(0.001)	(0.002)	(0.003)
(7) Security	−0.001	−0.001	−0.000	−0.001	−0.001	0.003	−0.003
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.005)	(0.003)
(8) People	0.003	0.001	0.001	−0.000	**−0.003*****	0.001	0.004
	(0.003)	(0.002)	(0.003)	(0.001)	(0.001)	(0.005)	(0.005)
(9) Investment	0.000	**0.004*****	0.000	−0.001	−0.000	0.005	0.000
	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)
(10) Profitability	**−0.008*****	−0.002	**−0.008****	−0.002	0.002	−0.001	0.009
	(0.003)	(0.002)	(0.004)	(0.002)	(0.002)	(0.004)	(0.006)
(11) FinancialService	−0.002	−0.001	0.001	0.001	0.001	0.008	−0.004
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.005)	(0.004)
(12) Health	−0.001	−0.002	**−0.002***	0.000	−0.000	0.001	−0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.008)	(0.003)
(13) Gaming	−0.002	0.000	−0.001	−0.001	0.000	−0.000	0.004
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.004)	(0.003)
(14) PlatformDevelopment	−0.002	−0.001	−0.001	−0.000	0.000	0.005	−0.001
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.005)	(0.003)
(15) Risk	**−0.006****	−0.003	**−0.006*****	−0.000	0.000	0.009	0.000
	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.010)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects (quarter-year)	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Inverse Mill's Ratio	No	No	No	Yes	Yes	Yes	Yes
Num. obs.	2505	1203	2403	369	369	345	466
McFadden/Adj./Nagelkerke R^2	0.254	0.190	0.112	−0.063	0.150	0.021	0.162

Note: The table provides the results for the sensitivity analysis with 15 topics derived from a separate sentLDA topic modeling results with K set at 15. The table includes results for models concerning ICO success and post-ICO performance, where the first three columns relate to models for *TokenTraded*, *Amount*, and *TimeToListing*, and the following four columns relate to models concerning *InitialReturns*, *InitialVolatility*, *LongTermReturns*, and *Delisted*. In Panel A, we report the results from the comparative out-of-sample tests of the prediction models with topic variables (Eq. 2) and without (Eq. 1). The performance metrics AUC, R^2, and concordance index (for the logistic, linear, and hazard models, respectively) are produced using simulated random data bootstrapped with 1,000 replications. The statistical significance of the differences in test statistics is determined with non-parametric Wald tests. In Panel B, the estimated coefficients concerning the relationships between the various white paper topics and ICO success and post-ICO performance measures are provided. The color-coded boxes indicate the estimated significance and direction of each topic variable's coefficients, where a green box indicates a positive relationship, a red box indicates a negative relationship, and a grey box indicates no significant relationship. As goodness-of-fit measures, McFadden, Adjusted, and Nagelkerke R^2 are provided for the Logit, OLS, and Cox P. Hazard models, respectively. *, ** and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on two-sided t-tests. All variables are defined in Tables 1 and 4.

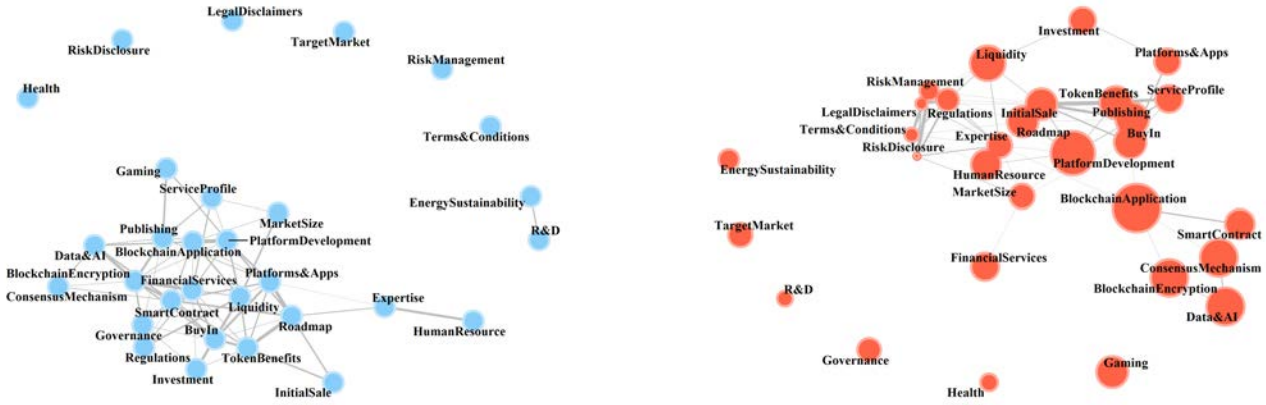
Table 10.: **White paper topics and ICO outcome – time analysis**

	ICO Success			**Post-ICO Performance**			
	TokenTraded Logit Model 1	(log) *Amount* OLS Model 2	*TimeToListing* Cox P. Hazard Model 3	*InitialReturns* OLS (2nd Stage) Model 4	*InitialVolatility* OLS (2nd Stage) Model 5	*LongTermReturns* OLS (2nd Stage) Model 6	*Delisted* Logit (2nd Stage) Model 7
Part A: pre-2018							
Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.716	0.115	0.743	0.069	0.103	0.105	0.529
Models with Topic Variables (Eqn 2)	0.717	0.119	0.760	0.067	0.039	0.019	0.568
Difference	0.001	**0.004****	**0.017*****	-0.002	**−0.064*****	**−0.086*****	**0.049*****
Significant Coefficients	Network **−0.005***	Blockchain **0.006***** Risk **0.008****		Innovation **0.005*****	Profitability **−0.002****		Mining **−0.013**** ICO **0.015***** Security **−0.017****
Num. obs.	316	207	290	90	90	84	115
McFadden/Adj./Nagelkerke R^2	0.249	0.208	0.303	0.024	0.128	0.043	0.275
Part B: 2018							
Model Comparison							
Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.735	0.069	0.730	0.005	0.028	0.012	0.537
Models with Topic Variables (Eqn 2)	0.741	0.076	0.746	0.005	0.014	0.014	0.537
Difference	**0.006*****	**0.007*****	**0.016*****	0.000	**−0.14*****	0.002	0.000
Significant Coefficients	Network **−0.002*** Blockchain **0.005**** Risk **−0.005*** Mining **0.003**** Innovation **0.002****	Network **−0.003**** Blockchain **0.007***** Mining **0.002***	Product **−0.002**** Blockchain **0.005***** Risk **−0.006**** Mining **0.002*** Innovation **0.002****		Security **−0.002****	Mining **0.006***	Mining **−0.005****
Num. obs.	1870	910	1797	251	251	236	320
McFadden/Adj./Nagelkerke R^2	0.161	0.119	0.141	−0.030	0.041	0.073	0.088
Part C: post-2018							
Model Comparison	*AUC*	R^2	*Concordance*	R^2	R^2	R^2	*AUC*
Base Model (Eqn 1)	0.569	0.060	0.798	0.430	0.101	0.182	0.509
Models with Topic Variables (Eqn 2)	0.566	0.063	0.833	0.385	0.152	0.237	0.503
Difference	-0.003	**0.003***	**0.035*****	**−0.045*****	**−0.051*****	**−0.055*****	**−0.006*****
Significant Coefficients		Risk **−0.018*** Profitability **0.007***	Network **−0.016*** Blockchain **0.013***** Risk **−0.015****				
Num. obs.	319	86	316	28	28	25	31
McFadden/Adj./Nagelkerke R^2	0.259	0.095	0.184	−0.017	−0.184	−0.018	0.000

Note: This table presents the results pertaining to time-specific analysis. Parts A, B and C of this table present the results obtained from sub-samples of ICOs issued in pre-2018, 2018 and post-2018 periods, respectively. Each part provides results for models concerning ICO success (*TokenTraded*, *Amount* and *TimeToListing*) and post-ICO performance (*InitialReturns*, *InitialVolatility*, *LongTermReturns* and *Delisted*). In Panel A, we report the results from the comparative out-of-sample tests of the prediction models with topic category variables (Eq. 2) and without (Eq. 1). The performance metrics AUC, R^2 and concordance index (for the logistic, linear and hazard models, respectively) are produced using simulated random data bootstrapped with 1,000 replications. The statistical significance of the differences in test statistics is determined with non-parametric Wald tests. In Panel B, the estimated coefficients for the relationships between the significant white paper topic categories and ICO success and post-ICO performance measures are provided. The color-coded boxes indicate the direction of the relat, where the green box represents positive relationships and the red box represents negative relationships. *, ** and *** denote statistical significance at the 10 percent, 5 percent and 1 percent levels, respectively, based on two-sided t-tests. All models include the specified control variables, which are defined in Tables 1.

8. Figures

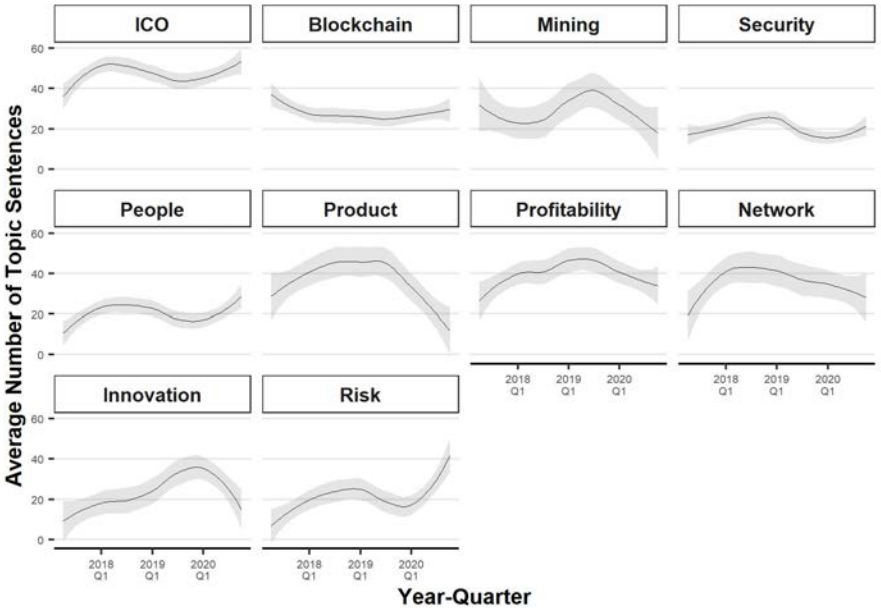
Figure 1.: **Network graphs on topic interlinkages**



(a) Word correlation among topics (b) Co-occurrence of topics across documents

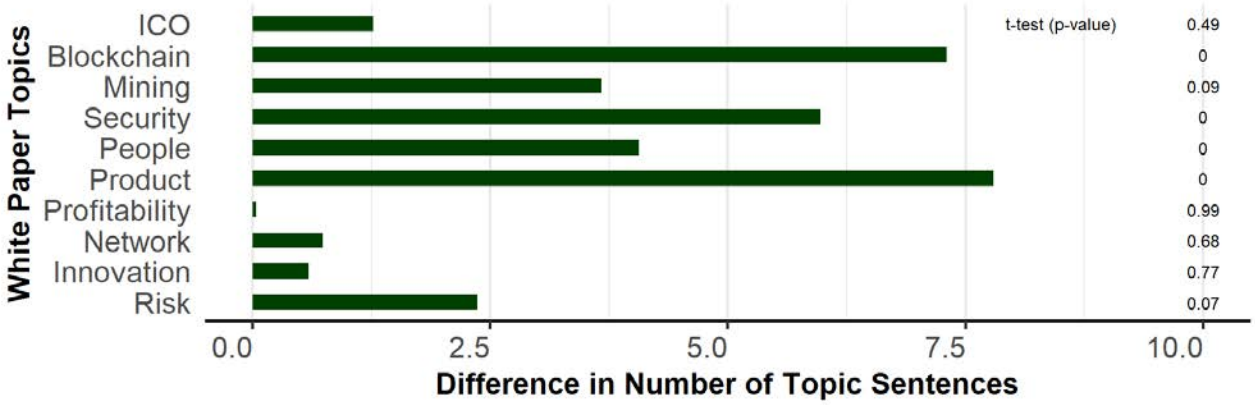
Note: Figure 1(a) presents a network graph illustrating the correlation between the white paper topics with respect to the weights assigned to the constituting words. Figure 1(b) presents a network graph illustrating the correlation between the white paper topics with respect to their co-occurrence across documents. In Figure 1(b), the size of the nodes proportioned to represent the number of sentences assigned to the given topic in the total sample.

Figure 2.: **White paper topic structure timeline**



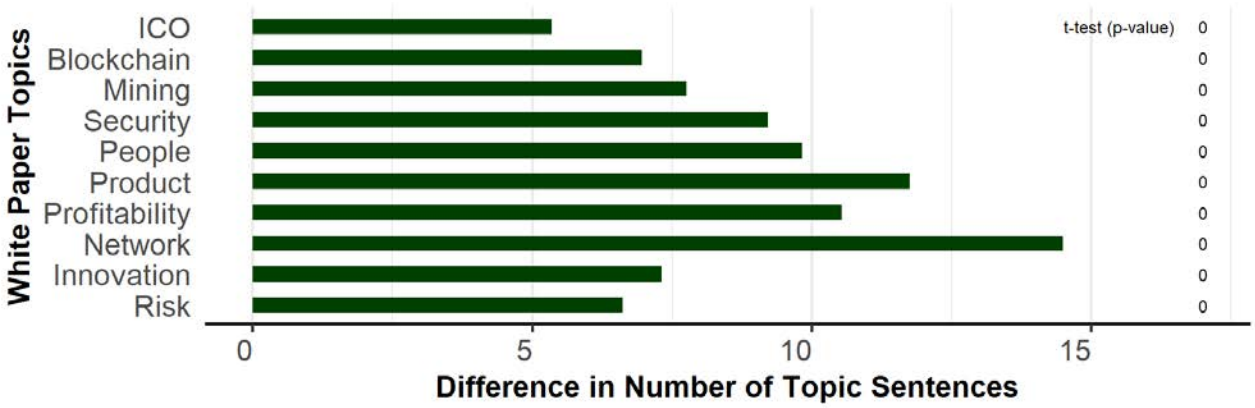
Note: Figure 2 presents smoothed timelines of the average number of sentences dedicated to each topic in ICO white papers in every quarter between 2017 Q2 and 2021 Q4 (quarters with more than 10 observations).

Figure 3.: **Average white paper topic sentences among regulated and unregulated countries**



Note: The figure presents the difference in the average number of topic category sentences in ICO white papers based in countries with ICO-specific regulations and those based in countries without clear guidelines on ICOs. The figure also indicates the p-values from Welch Two Sample T-Tests evaluating the differences. There are 1,139 ICOs from *Regulated* countries and 1331 ICOs from *Unregulated*. In aggregate, the average white paper length among regulated and unregulated ICOs is 336.25 and 299.57 sentences, respectively.

Figure 4.: **Average white paper topic sentences among high- and low- rated ICOs**



Note: The figure presents the difference in the average number of topic category sentences in white papers of ICOs with high ($\geq$median) and low ($<$median) aggregate ratings. The figure also indicates the p-values from Welch Two Sample T-Tests evaluating the differences. In aggregate, the average white paper length among high- and low-rated ICOs is 354.23 and 278.83 sentences, respectively.